UNIVERSIDAD DE CHILE
FACULTAD DE FILOSOFÍA Y HUMANIDADES
DEPARTAMENTO DE LINGUÍSTICA

# AN ANALYSIS OF VALIDITY OF A RUBRIC IN AN ACROSS THE CURRICULUM PROGRAMME OF ENGLISH AS A FOREIGN LANGUAGE

AUTORES: MARIA JESUS ALVARADO SANCHEZ, CHRISTIAN LAZCANO ESTAY, FERNANDA MORENO CASTRO, CONSUELO PEREZ ALARCON, MACARENA SALINAS MUÑOZ, JORGE SOTO ROJAS, CARLOS TAGLE GARRIDO

Informe Final de Seminario de Grado para optar al grado de Licenciado en Lengua y Literatura Inglesas

.

Profesor guía: Daniel Muñoz Acevedo

SANTIAGO DE CHILE
DICIEMBRE 2018

## ACKNOWLEDGMENTS

Este estudio no podría haberse realizado sin la valiosa contribución de Ricardo Úbeda, gracias por la confianza, ayuda y disposición, por su tiempo y los datos proporcionados los cuales hicieron posible esta investigación. También queremos expresar nuestra gratitud hacia los estudiantes y evaluadores de la Universidad de Santiago de Chile por participar en nuestro estudio y a toda la gente que se vió involucrada en este estudio y que ayudó a que se lograra. Finalmente, nos gustaría agradecer a nuestro profesor guía, Daniel Muñoz, por su disposición a abrir este seminario para quienes consideraba como los "rezagados", por guiarnos a través de esta investigación, además por todo el conocimiento, la paciencia y buena onda. Gracias de antemano por las cartas de recomendación, el almuerzo y gracias al Enzo quien sabemos es en realidad el que corrige la tesis.

**Seminario de Grado 2018**

Antes de todo quiero agradecer enormemente a mis compañeros de seminario por haber hecho esto posible y que saliéramos victoriosos a pesar de todas las dificultades que tuvimos en este largo proceso. Mejor team o mejor team.
También quiero agradecer a mi mamá por su apoyo incondicional durante todos estos años, desde el primer día estuvo conmigo acompañándome hasta hoy, sin ella no creo haber podido lograr nada de esto. Quiero agradecer también a mis hermanos por haberme soportado tanto durante este año (sobre todo por mi poquito mal humor). Gracias por solo darme un abrazo sin preguntar que me pasaba. Al Maty por cocinarme y alimentarme cuando llegaba muerta de cansada de la u y a la Sayi por alegrarme con sus ocurrencias. Mención especial para Hugo, que a pesar de todo siempre ha estado presente, aunque no sea su obligación, ya sea escuchándome, dándome ánimos o entregándome sus sabios consejos. Sin usted tampoco hubiera podido llegar hasta donde estoy hoy.
Finalmente, agradecer a mis amigas de la vida, Silvia, Kotte, Yessy, Ailin y Javo por sacarme un poquito de lo agobiante que es la u, por siempre subirme el animo y acompañarme durante en este proceso, solo ellas saben lo que me ha costado llegar a este punto, así que infinitas gracias.
¡Por fin se acabó la tortura de Inglesa!

**Maria Jesus Alvarado**

En primer lugar, quisiera agradecer a mis compañeras y compañeros de seminario, sin cuyo esfuerzo la realización de esta tesis no hubiese sido posible. En segundo lugar, quiero agradecer a nuestros profesores y profesoras por compartir sus conocimientos con nosotros, especialmente al profe *Daniel* por guiarnos en esta investigación, por ser muy buena onda y por siempre estar a disposición. En tercer lugar, quisiera agradecer a mis padres, *Christian* y *Elena*, por su apoyo y su arduo trabajo para que nunca me faltase nada, y a hermana, Emilia, quién a su modo, siempre está preocupada por mi. Mencionar también a mi polola, *Yaritza*, quién siempre está ahí para mí, brindándome apoyo y consejos incondicionalmente. Finalmente, quiero agradecer a todos mis amigos y amigas más cercanos de la universidad, cuya compañía fue muy importante para tener una estancia tranquila, amena y muy divertida en este campus durante estos 5 años que, sin duda, han sido los mejores de mi vida; muchas gracias por su amistad y buena onda, y espero tenerlos siempre a mi lado.

**Christian Lazcano E.**

*A mis padres,* Agradezco enormemente la ayuda que me han brindado a lo largo de este proceso, gracias por guiarme hasta este punto dándome la libertad de elegir lo que quería hacer en la vida sin presiones y siempre motivando mi vocación y cada locura que quería hacer. Gracias por el cariño y amor que me han dado en este proceso, por soportar mis infinitas mañas y crisis sin quejas, con infinita paciencia y comprensión. Este logro es también de ustedes por los valores, amor y responsabilidad con la que me criaron.

*A mis hermanas,* gracias a las mejores hermanas de la vida, quienes han sido un pilar fundamental en mi proceso de formación y me han ayudado en cada paso. Gracias *Clarita* por siempre escuchar todos mis descargos de mis pleitos universitarios, por traer esa alegría y drama necesarios que la vida necesita y que aunque no comprendas mis cosas académicas siempre me apoyas y te sientes orgullosa de mis logros. Gracias *Vicky* por compartir y motivar la lectura en mi vida, por compartir nuestros gustos nerd y esas memorables salidas al cine. Gracias por el apoyo y comprensión en todo lo que he querido emprender sin juzgar.

*A mis sobrinos, Marco Antonio* gracias por alegrar mi vida con toda tu ternura de bebé, por prestarme tu net todo el año, esta tesis no hubiera sido lo mismo sin tu ayuda. Agradezco también a mi sobrino que viene en camino por traer esta alegría a nuestra vidas.

*A mi prima, Barbara* gracias por ser como una hermana más en mi vida que me llena de alegría y con la que he pasado todos los procesos, gracias por acompañarme en el camino.

*A mi pololo, Felipe* te agradezco inmensamente el amor y cariño con el que siempre me tratas, que incluso en los momentos más oscuros de este proceso me iluminaba. Gracias por escuchar una y otra vez todos mis dramas, por entender mi estrés universitario y soportar mis tonteras de niña. Gracias por ser tan comprensivo y apoyarme a lo largo de este difícil trabajo que fue la tesis. Amor infinito para ti por ser el mejor compañero de vida.

*A mis amigos*, gracias totales a mi secta, *Ámbar, Andrea, Camila e Iván,* que estuvieron ahí todo el año desde distintas partes de Chile siempre apañando, gracias por todos los dramas y pelambres necesarios. Reyes del universo, Miserables, Rosos. *Ivancho* gracias por pasar este año extra conmigo, tkm.

*A mi perrito,* mención honrosa para Agustincito por ser el mejor perro del universo, por siempre recibirme con amor infinito, escándalo, alegría y rasguños de amor.

*Equipo de seminario,* gracias al profesor Daniel Muñoz por guiarnos a lo largo de este proceso y por el conocimiento compartido. Agradezco a mis compañeros de seminario por emprender este proceso conmigo. Es necesaria una mención honrosa también para la *Consu,* gracias por estar siempre apañando a lo largo de todo el arduo trabajo que tomó construir esta investigación, eres la mejor partner que pude tener en esta tesis.

**Fernanda Moreno Castro**

En primer lugar me gustaría agradecerle a mi familia, especialmente a mi mamá que me ha apoyado en todo lo que he querido hacer desde el principio de mi vida, siempre creyó en mí y en mis capacidades, cuando me inscribió en un colegio, que a pesar que no era en el que yo quería estar, descubrí mi potencial y me llevó a conocer a grandes personas que me acompañan hasta el día de hoy. No puedo dejar de lado a mi hermana Rosario que ha sido mi gran apoyo, la que me ha escuchado quejarme de todo en estos 5 años en la Universidad y además me ha hecho reír cada vez que lo necesité y siempre me soportó cuando todo andaba mal o bien. A mí Lalito que trato de hacerme la vida más fácil en la Universidad, cocinándome, yéndome a dejar cuando podía y que me inspiró muchas veces a escoger el mundo de las humanidades. A mi Salvita, que a pesar de ser un adolescente mañoso me hizo sentir el cariño a su manera de alguna u otra forma. A mis abuelos Ita y Cototo que son lo máximo. Por último no puedo dejar a fuera a mi Kaycita que llegó en mi primer año de la U y que siempre fue la más contenta cada vez que llegaba a la casa moviéndome su colita.

Gracias también a mis amigas, tanto a las que hice en la Universidad y las que llevan siéndolo desde el colegio. La verdad es que la vida estos 5 años no hubiese sido lo mismo sin ustedes Divinas y Círculo de Prada, sin ese entendimiento, ese apañe y esa sabiduría creo que me sentiría perdida.

En último lugar me gustaría agradecerle a todo mi equipo de trabajo, que a pesar de todo logramos sacar adelante esta tesis, mención honrosa para la Feña que de verdad hizo que toda esta locura de hacer una tesis fuese más fácil y al Profesor Daniel que nos apoyó en todo este proceso con su conocimiento, disposición y buena onda.

**Consuelo Pérez Alarcón**

En primer lugar, mi eterna gratitud a mi familia por soportarme durante los años que duró el error más caro de mi existencia, mis padres, *Leontina* y *José*, mi hermana, *Javiera*, a mis abuelos, tíos, tías, primos y primas por su cariño, preocupación, paciencia y entender mis ausencias en estos años en que la universidad me absorbió la vida. A mis amigos de la carrera que fueron la mejor compañía y fundamentales en este proceso, y quienes me dieron recuerdos memorables que me mantuvieron a flote hasta el fin de esta carrera. A mis amigos de la vida por sus consejos y ánimos en los momentos más difíciles. Mi gratitud hacia los profesores que hicieron esta experiencia un poco más llevadera, y en especial al profesor Daniel Muñoz por guiarnos y apoyarnos en esta importante etapa y a mis compañeros de seminario por hacer de esta tortura algo menos terrible. Finalmente, gracias a mi madre que me dio aliento y ánimos estos últimos meses convencida de que lo lograría a pesar de mis constantes llantos rogando por la eutanasia y a mi abuelo, *Jorge*, por tener tanta fe.

**Macarena Salinas Muñoz.**

Agradezco cariñosamente a mis padres, *Gloria* y *Jorge*, a quienes les debo todo. A mi hermana, *Cecilia*, por su apoyo. Al *Fede*, por existir. A mis compañeras y compañeros de seminario, por haber realizado un increíble trabajo. A nuestro profesor guía, *Daniel Muñoz*, por su disposición y sabiduría. A las profesoras *Cecilia Garrido*, y *Pascuala Infante*, y al profesor *Roberto Pichihueche*, por ser increíbles docentes, pero mejores personas. A mi compadre, *Kurt*, y a todas aquellas personas que conocí gracias a la universidad y que hicieron más amena mi estadía como estudiante, porque la verdadera vida universitaria ocurre fuera, y no dentro del aula. A todas y todos les recordaré con cariño y afecto.

**Jorge Soto Rojas**

Quiero agradecer afectuosamente a todos aquellos que contribuyeron en este estudio y en esta licenciatura. Muchas gracias a mi mamá *Catia* por su entrega incondicional durante todos estos años, sin ella esto no hubiera sido posible. Agradezco a mi papá *Carlos* y a mis hermanos *José Manuel* y *Fernanda* por todo su cariño y apoyo a lo largo de esta carrera, al igual que al resto de mi familia. También le doy las gracias a mis amigos más cercanos, que siempre me han dado motivación para seguir adelante. Le doy las gracias a mis compañeros del equipo de seminario de grado por su indiscutible entrega y gran aporte, tuve suerte al tenerlos como compañeros. Agradezco a mis compañeros de universidad, quienes se han convertido en grandes amigos y sin duda han mejorado mucho mi experiencia en esta carrera, gracias a ustedes recordaré este lugar con mucho cariño. Finalmente, agradezco a todos los funcionarios y profesores con los que me crucé durante tantos años, especialmente al profesor *Daniel Muñoz* por su buena disposición y gran ayuda durante esta tesis.

**Carlos Tagle Garrido**

## ABSTRACT

This investigation reports on a study of the validation of a rubric for oral exams in a context of standardised evaluation at university level in Chile regarding three aspects of validity: construct validity, face validity and reliability. The study examined the rubric applied in an across the curriculum English programme at USACH. The data was collected from three different formation programmes: Ingeniería en informática, Pedagogía en Lenguaje, Enfermería. The focus of the investigation relies on understanding the elements that contribute to the three aspects of the rubric's validity. This is a mixed-method study that collected self-reports by relevant stakeholders (students and raters) by applying a survey focused on their perception of the three aspects of the rubric's validity. Additionally, the actual rubric and associated documents (course programmes, student scores, etc.) were also analysed by applying a matrix of aspects of the three aspects of validity. The self-reports and the rubric were analysed in search for evidence that the rubric reflected adequately the correct functioning of the theoretical constructs of the instrument (construct validity, face validity and reliability). By making preliminary observations, it was expected that perceptions of stakeholders regarding the aspects of validity of the common rubric would reveal low levels of construct validity and reliability, and consequently, of face validity. The results showed that the rubric, indeed, presented problems regarding construct validity and reliability, but surprisingly, not regarding face validity, exposing the positive appreciation of the rubric of both stakeholders. Further research suggests for this issue to be addressed explicitly by considering the observation of variables that could explain this positive evaluation.



**Keywords:** construct validity, face validity, reliability, validation, rubric, stakeholders.

# TABLE OF CONTENTS

# Chapter 1: Introduction

In recent decades, language assessment has become a major issue in education policy around the world, raising important questions about the functions, conceptualizations, and applications of assessment in respect to curricula, teaching, and learning (Cumming, 2009). Over the last years, there has been much interest concerning language assessment issues such as the validity of score-based interpretations and the nature of the constructs we want to assess or the ethics and professionalism in the way we develop and use language assessment (Bachman, 2013).

A central issue within the area of second language assessment is the validity of assessment instruments, most importantly tests and rubrics. According to Bachman (1990) "a test is a measurement instrument designed to elicit a specific sample of an individual's behaviour" (p. 20). A rubric can be defined as a document that articulates the expectations for an assignment by listing the criteria of what counts and describing levels of quality from excellent to poor (Andrade 2000; Stiggins 2001; Arter and Chappuis 2007). As assessment tools, both tests and rubrics can be characterised in terms of two properties: validity and reliability. According to Moskal and Leydens (2000), validity is the extent to which a test measures what it claims to measure and reliability is defined as the consistency of assessment scores.

Within the field of language assessment, test validity has been more academically explored than validation models for rubrics in second language assessment. Andrade and Reddy (2010) indicate that "the research reports little study of the validity of the rubrics used. [...] Important aspects of validity have not yet been addressed at all, including the need to establish the alignment between the criteria on the rubric and the content or subject being assessed (content validity); the facets of

the intended construct being evaluated (construct validity); and the appropriateness of generalisations to other, related activities (criterion validity).” (p. 445). According to these authors, further research is needed along with extended explanations of the methods used to established validity (p. 446). Within this context, the aspect that has been more investigated in assessment is the reliability in the application of rubrics, with a focus on guidelines to follow in order to ensure the reliability of rubrics. At the same time there is a lack of a standard model that would ensure the validity of rubrics, as opposed to the assurance of their reliability.

Over the last years, the importance of learning English has been growing steadily, since it is nowadays considered to be a fundamental tool for future professional careers and programs of students. According to the national curriculum, the main purpose of the English Language curriculum is to get the students to develop the necessary skills to use the English language as a tool, as much as to access information as to solve simple communicative situations in oral or written contexts (Ministerio de Educación, 2009). The main problem of the English curriculum is related to the students’ socioeconomic situation, considering that it is implemented in different ways according to each school or educational institution, bringing inequality between students at the moment of applying to, and eventually entering higher education. This inequity directly affects the validity of the tests and rubrics elaborated by the national curriculum of English as the language skills assessments may not be coherent with the level of proficiency of students.

It is for this reason that the inequality of competences among students is evident when they enter university, especially in the field of English language teaching, in which this disparity is much more conspicuous than in other subjects. Consequently, the majority of universities have tended towards the implementation of across the curriculum English programs, in order for their students to ideally achieve the same level of proficiency.

The purpose of this thesis is to report on a study that analysed and validated a rubric from an across the curriculum programme focused on the teaching of English

as a second language. From the results obtained in this process, we expected to observe the degree of validity of this rubric based on established criteria in the field of language assessment.

To this purpose, we decided to carry out an investigation in Universidad de Santiago de Chile (USACH), due to its public character, the large amount of programs and courses that it offers, its institutional duty to be accredited and render account on what is being done academically, and, in general, due to the contribution that this kind of institution makes to the public education and society itself. This institution counts with an across the curriculum programme focused on the learning of English. As part of the programme, students of USACH are constantly evaluated through rubrics in language assessment, which makes them a suitable case to study the validity and reliability of a rubric in the context of a standardised testing system.

It is also very important to consider that a great number of graduate students face standardised tests, such as TOEFL or IELTS, in case they want to apply to an M.A or Ph.D. here or in a foreign country. As a result, studying the validation of a rubric becomes relevant for us, since through the evaluations that universities perform, students are placed at a corresponding level of performance. Consequently, universities should provide the proper contents that will allow them to continue their future studies.

The study was also motivated by our experiences as undergraduate students of English Linguistics and Literature. In that role, we have faced many instances of assessment that have lacked validity and reliability. For example, many times the corresponding rubrics for different tasks such as essays and tests were not shown nor explained to the students, which ended up affecting our performance and grades. This kind of situation is relevant if we take into consideration that a simple mark coming from an invalid assessment system might trigger serious consequences in a student's academic and/or personal life, whether the assessment's stakes are high or low.

Consequences of low degrees of validity in the design of a rubric can indeed be serious. Passing or failing a course might imply several different effects on the

academic and personal realm of a student. For instance, if a student fails a course, there can be a myriad of decisions taken by the student's family. As a case in point, if a student fails a subject that is needed as a requisite for the following level of the subject, his/her family will have to spend more money in order to pay for another year of university or rent as well. If a student is graded with a certain mark in order to apply for a scholarship, his/her life can change drastically. As a result, a simple mark provided by an invalid assessment system may imply several academic, personal and financial decisions to be made. It is thus essential that rubrics, and every assessment system in general, must be valid in order to provide fair grades to students.

At a general level, rubrics can contribute positively to an assessment system by promoting reliability in the educational system, as they establish the criteria that should be followed by evaluators. But if rubrics are to be applied, they must be reliable. By studying the validity and reliability of rubrics, we hope to contribute to the improvement in the validity and reliability in evaluations and assessment systems in general and also inspire interest in other colleagues and academics to conduct research on rubric validity and reliability.

This thesis is organised as follows: Chapter 2 reviews the literature of rubrics, validity and reliability. This chapter considers the definition of a rubric and the uses and different types of rubrics present in the field of evaluation and assessment. Accordingly, the relevance of this assessment tool is acknowledged as well as its history throughout time. Finally, the characteristics and components of rubrics are described: the concept of validity in rubrics is explained by providing definitions, explaining their relevance, identifying sources of validity such as construct validity, face validity, and consequential validity and by providing a comparison between test validation and rubric validation. In the end, the concept of reliability is developed, establishing the important components of the study reported in this thesis. Chapter 3 provides the methodology used to collect and analyse the data and informs the criteria and procedures used for the analysis. Chapter 4 mainly focuses on the results and the discussion around the outcomes presented before. Finally, conclusions and

final suggestions for further research concerning our investigation are presented in Chapter 5.

# Chapter 2: Literature Review

The objective of the study reported in this thesis is to examine and evaluate the use of a rubric for oral exams in a context of standardised evaluation in an across the curriculum programme. The field of rubric design is framed within the areas of language assessment and language evaluation, which are mainly related to educational contexts. As any assessment tool, rubrics must demonstrate the properties of validity and reliability. These properties will be explained in depth throughout this chapter. The review in this Chapter is organised as follows: first, all the aspects to the rubric are discussed in section 2.1; secondly, the main features of validity of rubrics are explained in depth in section 2.2. Finally, in Section 2.3 we discuss the aspects related to the reliability of rubrics.

## 2.1 Rubrics

In this section, the main definitions of rubrics and their theoretical foundation is presented. Most of the concepts in this section are based on Stevens, Levi and Walvoord (2013) in-depth description of rubrics.

### 2.1.1 Definition

Even though there are many definitions of rubrics, the model proposed by Stevens, Levi and Walvoord (2013) explains in detail the general features of a rubric that are relevant to this study. According to Stevens, Levi and Walvoord,, "a rubric is, as its most basic, a scoring tool that lays out the specific expectations for an assignment" (p.1). Also, as assessment instruments, rubrics are complex and have different components since they "divide an assignment into its component parts and provide a detailed description of what constitutes acceptable or unacceptable levels of performance for each of those parts" (Stevens, Levi and Walvoord, 2013, p.1).

In a similar fashion, Bachman (1990) describes rubrics as documents that inform students how to proceed in their evaluations as “the rubric of the test consists of the facets that specify how test takers are expected to proceed in taking the test. These include the test organisation, time allocation, and instructions” (p.118).

In summary, a rubric is an element that is used in evaluation contexts to decree what the steps to successfully perform a task are and to rule the importance and relevance of the aspects of the task that are to be evaluated.

### 2.1.2 Relevance of rubrics

Regarding the relevance of rubrics, Stevens, Levi and Walvoord (2013) state that “rubrics can be used for grading a large variety of assignments and tasks: research papers, book critiques, discussion participation, laboratory reports, portfolios, group work, oral presentations, and more” (p.3). They also indicate that “rubrics save time, provide timely, meaningful feedback for students, and have the potential to become an effective part of the teaching and learning process” (p.17). This means that,, if used correctly, rubrics can be very useful and versatile instruments capable of working as a guideline in regard to how students should proceed while going through their tasks. In addition, rubrics can give the students the possibility to realise about their mistakes in order to improve the next time they face an instance of evaluation.

As Stevens, Levi and Walvoord (2013) point out, there are a lot of reasons that one can find to use rubrics, and these reasons do not only have with efficiently using time and sounding pedagogy, but also, with other important basic principles that are obligatorily required in any instance of evaluation: the presence of equity and fairness. This means that the use of rubrics is framed within the conventions of a society that promotes values and principles such as the ones just mentioned: equity and fairness.

Rubrics also, as Stevens, Levi and Walvoord indicate, help students prepare to use thorough feedback, considering that the rubric and therefore the grading

criteria are to be discussed in class, the students have a much better notion of what the details in the feedback mean. Even when raters make vast notes and comments and students actually read them, it can be expected that there are still gaps between what the raters comment and what the student understanding of expectations may be.

According to Stevens, Levi and Walvoord (2013), rubrics boost and prompt critical thinking since by taking a look at the format of a rubric, students may detect for themselves certain patterns of problems that are constantly appearing in their performances or growing progress and improvement in their work. These discoveries may imply satisfactory outcomes when using rubrics. "By encouraging students to think critically about their own learning, rubrics can inspire precisely the pattern of "self assessment and self-improvement" intrinsic to creating the kind of motivated, creative students that are wanted in classes" (Stevens, Levi and Walvoord, 2013, p.21).

Another important feature of rubrics that Stevens, Levi and Walvoord highlight is that they are tools that ease communication with others, since teachers, evaluators and raters, in general, are constantly teaching mutually with others, either with the staff of a university writing center, tutors or remedial teaching staff, or with every other professor that is part of those students' learning process. Nevertheless, those "others" in the academic teaching life are, most of the times, teaching assistants of some sort. Rubrics then, "allow us to communicate our goals and intentions to all these people, sometimes without us even being aware that communication is taking place" (Stevens, Levi and Walvoord, 2013, p.23).

Stevens, Levi and Walvoord (2013) also argue that "rubrics help evaluators refine their teaching methods as in the same way that keeping copies of individual student rubrics can allow us to pinpoint a student's continuing improvement or weaknesses over time, rubrics showing student development over time can also allow evaluators to gain a clearer view of teaching blind spots, omissions, and strengths" (p.25). If, for instance, 70% of the students are having poor results in a particular aspect of a particular subject, let us say the absence of the third singular person

flexion in a writing skills class, raters should become aware of this situation and start talking more to the students about why and how the presence of the flexion is very important and if missing, there may be problems at the moment of communication. If a pattern of issues regarding defective use of examples is observed, this can be recognized and rectified.

In addition, Stevens, Levi and Walvoord claim that "rubrics level the playing field by establishing and setting evaluation criteria in order for the examinees to be evaluated under the same conditions and rules, boosting fairness and impartiality. One issue that is specific to the classroom experience is that of "translation." (p.27). This statement does not make reference to the fact that some or many of the students in the classroom may have communication problems when writing, reading, listening or speaking English, but to the fact that even native speakers of English may not totally understand or interpret the kind of English used in academia,, if it is the case, since not all rubric require academic English. Rubrics can act as very effective and avant-garde translation devices in this new environment. They do not only make what teachers are talking about understandable for students, but they also help raters comprehend where and when their words are not being understood or misinterpreted.

The role of rubrics is primordial since they are to be trusted if applied in the right way as they boost fairness and impartiality. In addition, rubrics are very useful and versatile instruments as they can be used for grading a large variety of assignments and tasks; they save time, provide timely, meaningful feedback for students; they help students prepare to use thorough feedback; they boost and prompt critical thinking; they ease communication with others; they help evaluators refine their teaching methods and finally, they level the playing field by establishing and setting a determined evaluation criteria. Considering all this, it can be stated that rubrics are essential to the assessment system in a course of language.

### 2.1.3 Types of rubrics

There are several types of rubrics including holistic, analytical, general, and task-specific. In relation to language assessment, the types of rubrics that are used most frequently are the holistic and analytic rubrics. On the one hand, as Mertler (2001) indicates, holistic rubrics are practical for evaluating a performance on a task as a whole. It provides an overall assessment of communicative skills as all criteria of what counts are assessed as a single score. Holistic rubrics are frequently used by evaluators as they tend to be easier to score. However, holistic rubrics do not provide detailed information on student performance for each criterion of what counts as the levels of performance are treated as a whole.

As indicated by Stevens, Levi and Walvoord (2013), a holistic rubric or scoring guide rubric contains a description of the highest level of performance expected for each dimension followed by room for scoring and describing in a "Comments" column just how far the students has come toward achieving or not achieving that level. However, this type of rubric is more time time-consuming when applied as it requires considerable additional explanation in the form of written notes. Also, scoring rubrics are useful for setting the expectations for and assignment but also require more time for raters in order to provide proper feedback. Scoring guide rubrics allow for greater flexibility and the "personal touch" of raters, but the need to explain in writing where the student has failed to meet the highest levels of performance does increase the time it takes to grade using scoring guide rubrics (Stevens, Levi and Walvoord, 2013).

On the other hand, analytic rubrics are useful for evaluating students' performance or different communicative skills. This type of rubric consists of two or more dimensions and each criterion is assessed separately, using different descriptive ratings (Mertler, 2001). There is general agreement that this type of rubric is capable of providing more specific feedback to students than the holistic rubric. This is relevant because a more detailed feedback means that the student will eventually identify his or her weaknesses and strengths.s pointed out by Arter & McTighe

(2001) “analytic rubrics are better suited to judging complex performances (e.g., research process) involving several significant dimensions. As evaluation tools, they provide more specific information or feedback to students, parents and teachers about the strengths and weaknesses of a performance” (p 22).

### 2.1.4 Components of a rubric

Stevens, Levi and Walvoord point out that rubrics are mainly composed of four basic parts. The steps, mechanisms and processes embroiled in designing a rubric can and should vary enormously, but the basic format prevails. “In its simplest form, the rubric includes a task description (the assignment), a scale of some sort (levels of achievement, possibly in the form of grades), and descriptions of what constitutes each level of performance (specific feedback) all set out in a grid” (Stevens, Levi and Walvoord, 2013, p.5-6). These components are explained below.

*Task description*

The task description in an instance of evaluation is almost always initially framed by the instructor and involves a “performance” of some sort by the student. A task can be very versatile and take the form of a specific assignment, such as an essay, a paper, or a dissertation. A task can also cover overall behaviour aspects, such as the use of proper protocols depending on the place of the evaluation or participation. Most rubrics are expected to contain a descriptive title and underneath it, a task description.

*Scale*

Stevens, Levi and Walvoord refer to scale as an element that “describes how well or how poorly any given task has been performed or occupies yet another side of the grid to complete the rubric’s evaluative goal” (p.8). Although there is no evidence that proves that the type of language used in the descriptors of the rubric may have some sort of effect on the student when reading the rubric, Stevens, Levi and Walvoord state that “in a generic rubric, words such as mastery, partial mastery,

progressing, and emerging provide a more positive, active verb description of what is expected next from the student and also mitigate the potential shock of low marks in the lowest levels of the scale" (p.8). In any case, some raters might prefer neutral, impartial and noncommittal language to be used, such as "high level," "middle level," and "beginning level," whereas others might prefer just numbers or even grades.

Even though Stevens, Levi and Walvoord indicate that there is not a set formula for the number of levels that a rubric scale should have, most evaluators suggest and prefer to clearly describe the performance of a task at three up to five levels using a scale, five levels being high enough for a valid assessment tool. On the one hand, "the more levels there are, the more difficult it becomes to differentiate between them and to articulate precisely why one student's work falls into the scale level it does" (Stevens, Levi and Walvoord, 2013, p.8-9). The explanation for this is that having too many levels and their respective descriptions may provoke when grading, considering that human being's performances are already difficult to classify, and will be even more difficult if there are too many categories in which they can fall. On the other hand, more specific and precise levels make the test task clearer for the student and,, at the same time, reduce the amount of time needed for the rater to make a detailed and appropriate evaluative process. Most raters then consider that three levels are the optimum number to be put on a rubric scale since since it seems to be a reasonable number in which levels and descriptors can be grouped without the necessity of adding or subtracting anything else that might alter the evaluation of the performance of the task.

*Dimensions*

Stevens, Levi and Walvoord refer to dimensions as a breakdown of the skills/knowledge involved in the assignment. Assessing language proficiency must consider that there are many communicative skills to be evaluated; therefore, there are many language abilities that a rubric must include in order to evaluate. The dimensions of a rubric are supposed to organized the parts of the task as simply and

completely as possible. A rubric is also expected to clarify how the corresponding task can be broken down into different components and their importance within the evaluation process, such as their scoring value. In an oral exam, for instance, the most important aspect to be evaluated may be the pronunciation or the organization of ideas as well as vocabulary or grammar. Furthermore, it is fundamental to be aware of how much weight is endowed to each of these aspects of the assignment since, for example, adding certain percentages to certain dimensions, accentuates the relevance of each aspect of the task. It is clear that if a student is familiarized with the rubric, he or she will know how his/her performance is going to be evaluated.

Dimensions also provide valuable feedback after the student is assessed. a language student has difficulties with pronunciation but no problems with grammar, then it should be noticed in the assessment provided by the rubric. The dimensions of rubrics must represent the type of component that students combine when being assessed in a test, such as examples, language appropriacy or content. Also, they must provide a quick examination of the students' strengths and weaknesses in each dimension. Also, the dimensions of a well designed rubric must be clear and free of any kind of value judgement. As Stevens, Levi and Walvoord point out, dimensions need not and should not include any description of the quality of the performance. "Pronunciation," for instance, is a frequent dimension in rubrics, but not "Good Pronunciation".

*Description of the dimension*

The dimensions of a rubic are complex and need a fair amount of work in order to be properly composed. As students perform in different levels of quality, the corresponding feedback of a language test must be precise and the dimensions of a rubric should describe the different levels of the expected performance. "Dimensions alone are all-encompassing categories, so for each of the dimensions in a rubric, a rubric should also contain at the very least a description of the highest level of performance in that dimension". (Stevens, Levi and Walvoord, 2013, p.10).

## 2.2 Validity of Rubrics

### 2.2.1 Defining validity

Validity as a key property of assessment referring the extent to which a test matches with the aim of the course or unit that is tested. Otherwise, the validity of a specific assessment refers to its capacity to determine if students have accomplished an established set of goals or level of competence (Brown, 2003). As such, validity is a property that can be extended to all assessment instruments, such as rubrics. According to Jonsson and Svingby (2007), validity must answer the following question: "Does the assessment measure what it was intended to measure?" (p. 136). The answer to this question, according to the authors, could be answered from two points of view, depending on the area in which the term is involved. For instance, if validity is seen as a test score interpretation, then it will be a property of the tests or interpretation of the test's results.

Another important feature of validity that Messick (1989) emphasizes is that it is both a unitary as well as a multifaceted phenomenon. This means that validity supports the inferences and interpretations from test scores and is also focused on the different types of evidence that need to be collected in order to support these inferences. According to Messick, it is important to highlight that validity is not only a function of the content and procedure of a specific test, but also involves how the tests takers perform at different instances of assessment.

In summary, the different definitions provided about validity in this section imply that the different characteristics of an assessment process can contribute to its general validity. The next section introduces and explains different sources and types of validity in order to have a grasp at the complexity of the construct.

### 2.2.2 Types of validity

Having established that validity is a unitary concept, there are other authors that state that this term can be classified into different kinds including: content, criterion and construct validity (Bachman, 1990). Content validity exists when content, relevance, knowledge and skills are revealed by the assessment (Jonsson and Svingby, 2007). According to Bachman (1990), criterion validity has to do with the relationship between tests scores and some criterion that could be an indicator of the ability tested. Finally, construct validity can be observed when asking: "does this test actually tap into the theoretical construct as it has been defined?". This term construct is thus related to a theory, hypothesis or model that intends to describe and explain a specific phenomenon of our world. (Brown, 2003).

Messick (1996) also claims that construct validity comprises six different properties of assessment, which are: content, generalizability, external, structural, substantive and consequential validity. On the one hand, generalizability means the extent to which score interpretations generalize across differents groups, occasions, and tasks. On the other hand, external aspect has to do with the relationship of the assessment score to other measures relevant to the construct being assessed. In relation to substantive validity that comprises theoretical and empirical evidence that reflect the thinking processes used by experts in the field (Jonsson and Svingby, 2007). Consequential validity has been described as one of the aspects of construct validity that support the validation of an assessment by describing all the implications of a test, including some considerations regarding the accuracy in measuring intended criteria, the impact on the preparation of test-takers, and the effect on the learners and the social consequences of the test's interpretation and use (Brown, 2003). Another important feature of consequential validity consists of the effect of the assessment on students' motivation and their performance in a course, their habits, independent learning and attitudes (Gronlund, 1998).

Another important aspect of consequential validity is face validity, which is related to how students perceive an assessment as fair and useful for their learning. This aspect of validation has been defined as the degree to which a test looks right and appears to measure the ability that is intended to measure (Mousavi, 2002). Face validity suggests how students notice the test to be valid. According to Brown (2003), face validity asks the question: "does the test, on the "face" of it, appear from the learner's perspective to test what it is designed to test?". There are many elements that help face validity to increase significantly. For instance, learners must find a well-constructed or expected format task, items that are clear, a task related to their coursework and a reasonably difficult level of the task.

Face validity cannot be empirically tested by an expert or a teacher, for the main reason that it is a factor of the "eye of the beholder" (Brown, 2003). This means that it depends on how the test-taker or the test-giver perceives the instrument.

The three types of validity explained in this section constitute properties of any assessment instrument, including rubrics. The next section introduces the idea of validity as observed in rubrics.

### 2.2.3 Rubric Validity

There are three types of evidence that are examined to test the quality of a rubric and determine if it measures what it is intended to measure: content, criterion and construct validity. These types of evidence are deeply related to two aspects of the rubric that are central for its validity, namely: the language that is used and the appropriateness of the rubric for the students that are going to be evaluated. Both aspects are connected to the types of validity since they can affect its interpretation and application. As Payne (2003) states, the appropriateness of rubrics is a matter of validity.

Clarity is also an important element because an ambiguous rubric cannot be accurately or consistently interpreted. It is important that the language used in a rubric is appropriate for the students to understand and for the instructors or scorers

to apply. Otherwise, this could lead to misleading that would affect the application and interpretation of the rubric. For instance, Moni, Berwick, and Moni (2005) gathered physiological concepts from dentistry students and faculty members in order to develop an assessment rubric that would be more accessible to the students and to more clearly represent expectations of each criterion and standard.

The content of the rubric is a key element for its validity as its dimension and descriptors must be adapted for the particular population of students that are going to be evaluated with this tool. Green and Browser (2006) tested transferring a rubric from the master's thesis literature of Shenandoah (SU) University to a similar programme at Best Practices University (BPU). There were considerable differences in scores among the raters of the two institutions because the students elaborated their literature reviews in different stages of their programmes. In addition, SU's students organize their projects according to personal and professional interests while BPU's students must include educational best practices into their projects. This is a clear example that rubrics must be created specifically to the body of students that are going to be evaluated. Adapting or using the one from another programme, even when they shared similarities, can lead to an imprecise evaluation.

It is important to point out that there are a number of issues related to the research of rubric validity. As stated by Andrade (2010), the studies available are not enough and a considerable amount of them does not provide complete information, some of them lack robust methodologies and, others do not provide results or a description of the procedures. Also, as Reddy and Andrade (2010) attest, the majority of the studies has been carried out by the authors in their own classrooms and there is a lack of replication that would contribute to the field and strengthen the results.

At the same time, research on this topic has been done mainly in the United States. This is especially relevant considering that every country has its own particular educational system, approach and evaluations that are adapted to their needs, culture and objectives. As Green and Browser's (2006) research proves, rubrics must be specifically designed according to the students and the objectives of a

class, taking into consideration every aspect of the evaluation because minimum elements can alter any of the elements of the rubric. There is not one rubric that can be used to assess everything, but rubrics are tools that must be created taking into consideration different variables. Some of these variables are general, like the programme and students, others more particulars, for instance, if its a written test, a report or a presentation or if the evaluation is one at the beginning or at the end of the programme. The aspects or skills that are going to be evaluated or even the date of the evaluation can add or subtract dimensions to the rubric. Applying the methods and results obtained in studies done in the United States can lead to imprecisions due to the different nature of the educational systems and their corresponding learning expectations.

Lastly, there is not a model to test the validation of a rubric. In the available literature there is not a definitive method to establish the validity of a rubric, and in the case of the studies that achieve it, it is not explained how they reached their validity. Moreover, the few studies that tested the validation of rubrics are focused on content validity, leaving aside other types of validity, like face or construct validity.

In summary, more research is needed on the subject of validation of rubrics in order to describe and explain the methods used to achieve such validation and to improve the tools for language assessment. The fact that the research on this topics is limited to the United States presents a problem considering that assessment expectations are specific to local educational systems..

### 2.2.4 Rubric validity vs test validity

Several authors (for example, Borsboom et al., 2004; McMillan, 2004; Messick, 1996) indicate that validity in educational research is often seen as involving evaluative judgment and is, therefore, not seen as a property of the test as such, but rather as an interpretation of the results. Thus, there is always a possibility that a test could be biased when systematic differences in test performance occur that appear to be associated with characteristics not logically related to the ability in

question (Bachman, 1990). The implementation of a rubric in these cases would then be adequate in order to ensure a more valid assessment process.

Nevertheless, it is important to notice the presence of bias in rubrics as well. This happens when a rubric is not validated in its different aspects. For instance, when a rubric has been shown to have content validity, and no other aspect of validity has been addressed, it could mean that content knowledge is properly assessed, while other dimensions, like thinking processes, are not. It could also mean that there is no alignment between objectives and assessment, or that there are severe social consequences or bias. All these factors threaten validity and might produce unfair results, in the sense that students are disadvantaged in their opportunity to show what they have learned. (Jonsson and Svingby, 2007). Since the validity of tests does not seem to be an inherent feature of tests, it is important to note that the implementation of a rubric would not necessary solve validity issues, since lack of validity can occur in rubrics as well.

Considering that validity, as an aspect of language assessment, has been described mostly in tests while rubric validity is still lacking development, it can be argued that scoring with a rubric does not necessarily provide higher levels of validity than scoring without one. Merely providing a rubric does not necessarily provide content representativeness, fidelity of scoring structure to the construct domain or generalizability. Nor does it give any convergent or discriminant evidence to other measures. There is, however, one aspect of validity that might benefit from the use of rubrics. If rubrics in some way affect instruction, so that there are positive educational consequences from using them, then this would influence the aspect of consequential validity (Jonsson and Svingby, 2007). According to Messick (1996), this aspect of validity includes evidence of implications of score interpretation, both intended and unintended as well as short- and long-term consequences.

The implementation of validation models in testing systems has been rather new, especially those related to rubrics. Considering that tests as well as rubrics are not inherently valid, the implementation of both simultaneously would at least

contribute to provide a minimum level of validity, even if it is in some aspect(s) of it, like face or consequential validity.

## 2.3 Rubrics and reliability

### 2.3.1 Defining reliability

Brown (2004) defines the concept of reliability as a consistent result of a test, in the sense that, if a student takes the same test two times, he or she should obtain the same or a similar result for the same test. If this is not fulfilled, the test would not be reliable (p. 20). Luoma (2004) adds that "reliability is important because it means that the scores are dependable, so that we can rely on them in decision-making" (p. 176). According to Bachman (1990) "when we increase the reliability of our measures, we are also satisfying a necessary condition for validity: in order for a test score to be valid, it must be reliable" (p. 160). This means that both concepts are causally related to each other. In other words, a test score cannot be valid, if it is not reliable.

There are several factors that can affect the consistency of results. One of the most important factors is, according to Bachman (1990), the communicative language ability of the learner, which consists of language competence, strategic competence, and psychophysiological mechanism in communicative language use (p.14). These factors are, of course, up to the test taker responsibility only. However, there are other factors that affect language test scores which go beyond the individual learner. Bachman (1990) describes three as the main factors: (1) test method facets, (2) attributes of the test taker that are not considered part of the language abilities we want to measure, and (3) random factors that are largely unpredictable and temporary.

Test method facets are systematic to the extent that they are uniform from one test administration to the next. We can understand facets as features of the evaluation instrument; such as the design and duration of the test, the different types of

exercises, the abilities that the test is measuring, the previous preparation that the test needs, etc. Bachman (1990) explains that, if the input format facet is multiple-choice, this facet will not vary whether the test is given in the morning or afternoon, or whether the test is intended to measure grammatical or illocutionary competence (p. 164).

For example, there are certain tests that need a previous preparation to be taken such as CPE (Certification of Proficiency in English), which requires that a course is taken before taking it. On the other hand, there are tests such as TOEFL (Test of English as a Foreign Language) that may not require any prior preparation. In the first case, if the student takes the test without any prior preparation, his/her results will be affected, because they did not have any knowledge about the instrument and not because of his/her language ability.

Attributes of the test taker that are not related to language ability are also systematic, in the sense that they are likely to affect a given individual's test performance regularly. Bachman (1990) exemplifies it as follows: if the design of the test or the format of the same is not suitable for the students' ability, the student will not be able to achieve properly the objective of the test, even if they actually have the knowledge required. If the design or format of the test affects the individual's performance on one cloze test, it is likely to affect his performance on any cloze test (p. 164), because the characteristics of the test are beyond the communicative language ability of the test taker.

Random factors, unlike the other two factors explained above, are unsystematic. According to Bachman (1990), these include unpredictable and largely temporary conditions, such as mental alertness or emotional states, and uncontrolled differences in test method facets, such as changes in the test environment from one day to the next, or in the way different test administrators carry out their responsibilities (p. 164).

It is important to highlight that, according to Bachman (1990), the major concern in the design and development of language tests is to minimize the effects of

test method, personal attributes that are not part of language ability, and random factors on test performance (p. 167). Therefore, the existence of a rubric becomes a crucial aspect in language assessment, in order to reduce the implications of other factors that are not related to the language ability of the students. Rubrics should be explicit in its dimensions and descriptors according to language ability and to the students' ability to perform certain tasks.

### 2.3.2 Types of reliability

There are two types of reliability: intra-rater reliability and inter-rater reliability. As stated in Luoma (2004), "[...] intra-rater reliability or internal consistency, [...] means that raters agree with themselves, over a period of a few days, about the ratings that they give" (p. 179). Along these lines, Brown, Bull, and Pendlebury (1997) state that the "major threat to reliability is the lack of consistency of an individual marker" (p. 235). Along the same lines, Jonsson and Svingby (2007) pointed out that, regarding intra-rater reliability, there is no concrete evidence that shows a lack of consistency within raters when they are using a rubric. In contrast, raters using rubrics as assessment instruments showed sufficient consistency when it came to assessment in the educational field.

Another type of reliability that Luoma mentions is inter- rater reliability, "which means that different raters rate performances similarly. They do not necessarily need to agree completely [...] However, if the raters use the same criteria, their ratings should not be wildly different" (p. 179-180). In this sense, inter-rater reliability presents more challenges than intra-rater reliability. This is because the data related to intra-rater reliability is more difficult to observe than inter-rater reliability. In this sense, research done within this area demonstrates a lack of exact agreement between raters when it comes to assess with assessment instruments.

According to Jonsson and Svingby (2007), there are two main methods of measurement estimates: the Rasch model and the Generalizability theory. These models are different correlation coefficients. These models suggest that the

satisfactory percentage of raters' agreement should be 70% or greater in order to achieve reliability in an assessment instrument. Brown (2004) also argues that consistency estimate values above 70 are acceptable. Nevertheless, in most studies, the generalizability theory has been the most used regarding the matter of methods of measurement regarding reliability.

Bachman (1990) and Brown (2004) also agree that we must bear in mind 'the true score measurement', in order to understand that what raters really observe of their students' competence is not always aligned with the real competence that exists in the students' mind. As Bachman (1990) points out, language abilities are abstract and thus test takers can never directly observe them. The true score would be observable if we had access to the student's mind (p. 166). As achieving the true score measurement is impossible, we must trust that the test is as reliable as possible, in order to obtain the results that most resemble the real language ability of the students. Brown (2004) proposes four types of reliability characterising the main factors that affect the 'true score' of test takers: (1) student –related reliability, (2) rater reliability, (3) test administration reliability and (4) test reliability.

Student- –related reliability is related to all the factors associated to the state of mind of the test taker, such as fatigue, anxiety, or any other physical or psychological factor, which may affect the communicative language ability of the student. If the test taker is suffering one of these inconveniences, the observation of the true score will be affected, because the test taker performance is being affected as well and, consequently, the degree of reliability of the results.

Rater reliability is associated with the concept of inter-rater reliability, because human error, subjectivity, and bias may enter into the scoring process. That is, if one test has two evaluators who differ from each other very widely, the process of evaluation will be considered unreliable. That may be the consequence of lack of attention to scoring criteria, inexperience, and inattention or even preconceived biases; all such factors causing those two scorers are not using the same standard.

On the other hand, lack of agreement may occur even when there is only one scorer. Brown (2004) explains that an evaluator who has to score 40 tests runs the risk of being inconsistent in its evaluation. It is likely that the first 10 tests will be evaluated under different criteria than the last 10, for example. This is due to different states that scorer can experiment through the development of the evaluation process, such as fatigue or lack of concentration. In this particular point, the relevance of the rubric appears, as the dimensions and descriptors that the rubric provides, help raters to compensate the possible bias that may affect the assessment by helping to ensure that these factors do not have a major impact on the results.

Test administration reliability refers to all external factors that may affect the student at the time of taking the test, in other words, and the conditions in which the test is administered. Brown (2004) exemplifies these external conditions as follows: if there is street noise during a listening activity and that affects the students' comprehension of the tape, the test administration becomes immediately unreliable. Other sources of unreliability are found in photocopying variations, the amount of light in different parts of the room, variations in temperature, and even conditions in chairs and desks (p. 22).

Test reliability is related to the test itself, so that if the test is too long, the results obtained by the students may be affected, because they may become exhausted by the time they reach the later items of the test, which, ,in turn, could affect if they respond correctly the test. This is why the design of a test and its tasks may be more appropriate for one type of student than for another.

In this sense, there are several factors that affect the reliability of an assessment. To reduce the bias that these factors may produce there are procedures for ensuring reliability, which are going to be explained in the following section.

### 2.3.3 Procedures for ensuring reliability

In contrast to validity, there are several procedures for ensuring reliability, especially among raters. Luoma (2004) defines how to achieve this

consistency at the moment of rating. He proposes that there should be a rater training before formal examinations standard setting, or the setting of cut scores or other standards of success. At the same time, , it is important to carry out a correlation study, in order to identify the reliability of the scores.

Luoma (2004) explains that rater training sessions usually last several days. In these sessions, the prospective raters are selected according to their interests in and their experience of testing. The main focus of the training sessions is to practice and have knowledge of all the scenarios that a rater may experience at the moment to evaluate real performances. In order to achieve the goals of the training, the examiners have to listen and discuss several tapes that contain different levels of performance. At the end of the course, the examiners usually have to go through a qualification procedure, which involves evaluating a recording to be compared with other qualified rater(s) in the system.

In the same manner, Luoma (2004) describes another procedure for ensuring reliability known as standard setting. In this process, examiners evaluate different levels of proficiency, in order to define the cut point between masters and non-masters. In this way, raters get a real knowledge about the different levels that test takers can achieve, having in mind the right and wrong answers that may appear during the assessment. Luoma (2004) mentions that "these procedures help test developers explain how the test scores are related to speaking skill outside the test" (p. 178). Therefore, this procedure of ensuring reliability serves to maintain the same standards through different versions of the test, especially in formal examinations.

According to Luoma "The most common way of expressing reliability is through correlation, which is a statistical indicator for the strength of relationship between variables." (p. 182). Correlation is a type of analysis of two sets of data ordered as scales which may be related or not. In this sense, there are different methods of correlations that are used to ensure reliability in language assessment. Spearman rank order correlation, Bachman correlation and the Pearson product–moment correlation are the most used correlation procedures concerning reliability.

The common aspect that these correlation tests is that they determine whether two sets of scores may increase or decrease in the same way and proportion.

When it comes to assessment "a rubric can be seen as a regulatory device for scoring, it seems safe to say that scoring with a rubric is probably more reliable than scoring without one." (Jonsson and Svingby (2007, p. 136). Even though there are certain guidelines that ensure the reliability of an assessment, they are focused on determining reliability among raters, the rater himself/herself and the test application. There are still no guidelines to ensure reliability of rubric as instruments of assessment that observe their design features.

### 2.3.4 Relevance of reliability for rubrics

Reliability is relevant because the assessment of intended learning outcomes must be consistent. The required consistency promotes similar scoring patterns, and can be ensured by training raters, and also the use of multiple raters in the assessment process.

In language assessment, it is well-known that rubrics enhance the consistency concerning scoring students, tasks and assignments. According to Jonsson and Svingby (2007), an important aspect of reliability concerning rubrics is that failing to accomplish reliability has direct consequences on the students assessed. For this reason, assessment instruments have to be credible and trustworthy independently of who the rater is or the place and time that the assessment was taken delivering similar results. In actual practice, reliability is extremely hard to achieve, because a rubric can be naturally interpreted and used differently by different raters.

As explained in section 2.3.3 above, the true score could only be observed if we had access to the test takers' mind. In consequence, raters will never reach the true score, but a real score. According to Brown (2004), rubrics aim to minimize the differences between these two types of score by increasing the level of reliability of the assessment. It is for that reason that the need to rely on a rubric arises, on account

of maximizing the level of reliability, but, at the same time, decrease the inconsistencies among raters and within raters as well.

In this chapter, the main aspects of validity regarding language assessment were described and explained, with special attention to the validity and reliability of rubrics. It has been shown here that there is still no model for ensuring validity of a rubric as a whole, considering construct validity, face validity and reliability. On the other hand, even though there are guidelines used for estimating reliability, such as correlations and the recognition of factors that may affect reliability, there are no guidelines that ensure the reliability for the rubric as an instrument of assessment. Additionally, there is a lack of research regarding rubric validity, and the already existing studies are usually biased or less representative given that researchers use their own students or courses to carry out their investigation. The available literature is also biased as it mainly represents the reality of US assessment.

The study reported in this thesis, addresses some of the gaps observed in this chapter and by analysing the validity of a rubric as an assessment instrument in an across the curriculum programme. In the next chapter, the methodology of this study will be addressed.

# Chapter 3: Methodology

The purpose of this investigation was to analyze three main dimensions of validity in a rubric: reliability, construct validity and face validity. The rubric examined is associated to an oral exam applied in an across the curriculum programme at a public university in Chile. In general terms, the methodology used in the study consisted of the analysis of the actual rubric and the application and analysis of a survey to students and raters involved in the English programme. The study was guided by the following research questions:

- RQ1: What is the degree of validity of a standardised rubric regarding construct validity, face validity and reliability?

  - RQ1.1: What are the raters' perceptions of validity regarding the three aspects?

  - RQ1.2: What are the students' perceptions of validity regarding the three aspects?

  - RQ1.3: What are the features of the rubric regarding the three aspects?

To answer these questions, we conducted a mixed-method study that integrated the application and analysis of surveys applied on raters and students and the analysis of the actual rubric used by the programme. In the first place, one part of the data was collected through two surveys: students' and raters' surveys that consisted of a series of questions presented in different formats: open, close-ended, rating scale, multiple choice, etc. The responses to the survey were analysed in search of the opinions and ideas participants had regarding the face validity, construct validity and reliability of the rubric. Simultaneously, the analysis of the oral exam

rubric was carried out by comparing the rubric with the actual course programme and by observing the descriptors and levels of the rubric.

## 3.1 Context of the study

The rubric analysed was part of the assessment of oral proficiency in the across the curriculum programme of English at Universidad de Santiago de Chile. Universidad de Santiago de Chile (USACH) is a public university in Chile with more than 21,000 undergraduates as of 2018. It has an across the curriculum programme of English called Blended Learning that consists of four courses aimed at proficiency levels, from A1 (ALTE Breakthrough) to B1 (ALTE 2). These levels take after the levels set by the Cambridge University to determine the proficiency of L2 learners of English.

The programme adopts blended and semi-presential class strategies, which is a combination of face-to-face and online classes. The class strategies put special emphasis on tutorships designed to improve students' communicative and linguistic skills such as speaking, reading, listening, and writing. Every course has different thematic units according to the level of difficulty that the different language aspects that are being taught imply, from daily life vocabulary to more complex and specific topics, such as those that involve technical English like medicine, engineering, economy, among others. The Blended Learning programme has four biannual modules equal to 25.5 in-person work hours and 64.5 distance work hours, and it also requires a minimum of 30 minutes of daily online classes per student.

The programme has three types of evaluation processes. The first is constituted by formative evaluations serving the purpose of showcasing the performance of students in relation to the specific objectives of units during the development of the course. The second type of evaluations is of summative nature and they are taken at the end of each teaching unit. These summative evaluations serve the purpose of showcasing the level of linguistic competence acquired in relation to the proposed objectives from each thematic unit. The third type of

evaluation consists of two oral evaluations: one mid-term, and one final exam. These two oral exams are assessed with the same rubric, which is the one examined in this study.

## 3.2 Participants

Participants of the study were 23 students between eighteen and twenty-one years old, who belonged to the programmes of Ingeniería en Informática, Pedagogía en Lenguaje, and Enfermería. Surveys were also applied to14 raters teaching in the English programme with at least 3 years of experience. They were surveyed regarding their history of elaborating, applying, and/or using the oral exam rubric.

We decided to select these three dissimilar programmes due to the fact that the importance of English is different among them, and thus, we expected more diverse responses according to the different profiles of students in terms of their motivation to learn English. On the one hand, students of Ingeniería en Informática need a high level of English proficiency because the area of programming and networking is a mostly developed in English. In a similar way, students of Enfermería also need a medium to high level of English proficiency in order to access to the scientific information they access, which is mostly written in English. On the other hand, students of Pedagogía en Lenguaje were not expected to require the same level of English as the other two programmes, because the English language is not a central concern of their profession. Still, they need a basic level of English proficiency in order to deal with certain circumstances of exposure to the English language.

## 3.3 Data collection

Three types of data were collected for the study. The first one was the responses obtained from the application of a survey focused on the evaluators' experience with rubrics. The second set of data consisted of the responses obtained by the students' experience with rubrics. The third data set consisted of the

observations made to the components of the rubric (dimensions, levels and descriptions) regarding the three dimensions of validity set in the research questions (construct validity, face validity and reliability).

### 3.3.1 Data collection tools

To answer RQ1 and RQ2, regarding teachers' and students' perceptions; construction and reliability of the rubric, two surveys were created. In order to answer RQ3, the oral exam rubric was used. The surveys were designed formulating roughly the same questions, to later compare the answers of both stakeholders (students and raters). The questions from the surveys were designed as described below, in order to observe them in accordance with the criteria for construct validity, face validity, and reliability of the rubric.

*Raters survey*

The survey designed for raters addressed four areas of interest:

*General information*

This section was composed by six questions oriented to elucidate the raters' experience with the rubric and as evaluators in oral exams. For example, "¿Cuantos años de experiencia tiene como evaluador?" and "¿Cuántos años de experiencia tiene como evaluador en este programa?".

*Oral exams' design and application*

This section consisted of eight questions, aimed at establishing the perceptions of the raters in relation to the criteria used in the design and application of the rubric of the oral exam. For instance "¿Cree que la rúbrica corresponde con lo que enseña/práctica en el curso y con lo que evalúa en los exámenes orales?".

*Evaluation*

In this section, the raters were asked about how they use the rubric, in relation to themselves and other raters; additionally, in order to clear up the consistency of the assessment among rater. For example, "¿Cómo entrena para utilizar la rúbrica de manera consistente?" and "¿A qué aspectos de la rúbrica le da más importancia al momento de evaluar? ¿Por qué?"

*Opinion*

This section was constituted by ten questions. The target was for the teachers to provide a general opinion of the weaknesses and strengths of oral tests as well as aspects they suggested to change in relation to the tests. For instance, "¿Qué aspectos de la habilidad oral de los estudiantes agregaría a la rúbrica?" (A copy of the survey has been included in Appendix A).

*Students survey*

Surveys for students contained questions that mirrored, in order to compare the perceptions of the stakeholders according to the rubric. The survey was designed to address the following areas:

*Personal information*

This section consisted of six questions. The aim of this section was to establish the experience of the students in relation to the oral exam rubric and programme in general terms. For instance, "¿Tienes acceso a la rúbrica antes de una evaluación?" and "¿Tienes acceso a la rúbrica antes de una evaluación?".

*Satisfaction Level*

This section was composed by eight items, involving the students' personal evaluation in the design and application of the oral exam rubric. For example: "¿Estás satisfecho con la rúbrica de los exámenes orales?".

*Knowledge*

This section consisted of four questions. Their main objective was to observe the knowledge the students had regarding the design and application of oral tests. For instance: "¿Tienes acceso a la rúbrica antes de una evaluación?" and "¿Los profesores te han explicado alguna rúbrica del examen oral?".

*Opinion*

In this section, ten questions were made, in order to elucidate the opinions that the students could have in relation to the rubric and programme. The aim of this section was for the students to provide their opinions of the weaknesses and strengths of the oral exam rubric, as well as the aspects they would change in relation to the oral exam rubric. For example, "¿Le cambiarías algo a la rúbrica del examen oral?, ¿Qué?" and "¿Crees que hay aspectos importantes de tu habilidad oral que no son evaluados en la rúbrica?".

The surveys were piloted with a focus on identifying issues related to the length and clarity of the surveys. The student's survey was piloted by eight random students from the Universidad de Santiago, who did not participate in the programmes of Enfermería, Ingeniería en informática, and Pedagogía en Lenguaje. The observations from these participants were obtained in relation to aspects such as length and clarity and were used to prepare the final version of the survey. In turn, the raters' survey was piloted by the thesis supervisor under the same criteria of length and clarity.

*Rubric*

Finally, to answer RQ3, we obtained the oral exam rubric with all the documentation of the English programme of USACH. The actual rubric corresponds to an analytic one and it was analysed by comparing its contents to being compared with each English course syllabus in the four levels of the programme and by observing its components according to the dimensions of validity under research.

The instrument analysed in this investigation corresponds to a rubric adapted from Cambridge, with the purpose of being used in an across the curriculum programme in Universidad de Santiago de Chile. This adapted oral exam rubric is organised according to five dimensions,namely Achievement, Linguistic Range, Organization, Accuracy and, Register and Academic Discourse. Each dimension points out to answer a particular assessment dimension for students' performance in the oral exam. For instance, in Achievement, the students have to complete the task successfully and without difficulty. In Linguistic Range, the students need to use a wide variety of vocabulary and grammar structures. Then, in *Organization*, if the ideas expressed clearly and connected effectively. Accuracy says if the grammar structures were produced correctly. Finally, Register and Academic Discourse observed if the student use language with an appropriate level of formality. Each dimension includes six levels/bands that are positioned in ascendant order of accomplishment, according to the student's performance. A copy of the rubric is included in Appendix E.

### 3.3.2Data collection procedures

Firstly, the thesis supervisor made the initial contact with the Programme Coordinator of the English Department of Universidad de Santiago de Chile and let him know our interests in working with him. The programme coordinator agreed to meet with the team to talk about all the details concerning the investigation. In this meeting, the Coordinator agreed to collaborate with the study and provided all the information associated to the programme of English was provided, including the actual rubric of the English programme that is currently used by the same.

The surveys were applied online through the Programme Coordinator, via mail. The data was collected for approximately two weeks. The responses that the surveys provided were collected as data for the study.

## 3.4 Data analysis procedures

The two surveys consisted of open-ended questions, close-ended questions and likert-scale questions. Responses were classified and analysed (see Appendix F) according to the different research questions and dimensions under examination (see Appendix C). To this purpose, responses to the surveys were tabulated as illustrated in the following table.

Table 16
*How do you train to use the rubric consistently?*

| Answers | Frequency |
|---|---|
| Self-training | 4 |
| Non-training | 3 |
| Off-topic | 3 |
| Rater training | 1 |

Responses to the surveys were interpreted in relation to the validity dimensions of face validity, construct validity and reliability (see Appendix C). Ideas identified in the responses to open-ended questions were interpreted as representing the participants' perspectives on the corresponding dimension of analyses. Responses were categorised by tagging content labels that allowed to compare amongst teachers and students. The list in Table 2, illustrates how content category tags were assigned to the content of responses.

Table 2
*How do you train to use the rubric consistently?*

| Answers | Tag |
|---|---|
| Leyendo los criterios y analizando casos posibles para cada uno<br>La reviso bien<br>La reviso muchas veces<br>La semana previa al examen vuelvo a leer la rúbrica para tener claridad sobre ella y sobre qué tipo de errores caería en cada categoría | Self training |

We also analysed the content of the across the curriculum programme in comparison to the oral exam rubric. To this purpose, we compared the rubric for oral exams and the content of the syllabi of each level of the programme (Levels I, II, III and IV). In addition, we identified validity issues in the main elements of the rubric: levels, dimensions, descriptors. For example, descriptors of the rubric were examined in search of keywords such as qualifiers in order to observe the construct validity of the rubric.

# Chapter 4: Results and Discussion

In this chapter, the results of the study and the discussion of the results are reported. The chapter is organised according to the research questions guiding it. Research questions include the three central aspects of this study: construct validity, face validity, and reliability. In this manner, the chapter begins with the findings for RQ1.1, aimed at discovering the raters' perceptions regarding the three aspects of validity. Results for RQ1.2 are reported next, indicating the students' perceptions of the three aspects of validity and RQ1.3 provided the features of the rubric regarding the construct validity. Finally, the discussion of the results is addressed in order to examine the degree of validity of the rubric.

## 4.1 Results for Research Question 1.1: Raters' perception about the rubric

Under RQ1.1, three sub-questions were posed indicating the perceptions of the raters regarding the three different aspects of the validity of the oral exam rubric: construct validity, face validity, and reliability. Results suggest interesting issues with raters' knowledge of the rubric, as discussed below.

### 4.1.1 Results for Research Question 1.1: Raters' perception regarding construct validity

In relation to raters' perceptions about construct validity, regarding the question illustrated in Table 3 below, the results indicate that most of the raters do not believe that the rubric is adapted to the programme, arguing that this is an across the curriculum programme while others do believe that the English programme adapts adequately to each programme. One rater was not sure about his/her answer, which affects in a direct way the construct validity of the rubric. According to the literature, there should be an alignment between the objectives and the programme (as indicated in section 2.2.4), so the view expressed by raters indicates a lack of

construct validity due to a lack of alignment between the rubric and the contents of the course.

Table 3

*Is the English programme adapted to the needs of each programme?*

| Answers | Frequency |
|---|---|
| Yes | 3 |
| No | 9 |
| Not sure | 1 |

When the raters were asked about the part of the oral exams that was most difficult to evaluate for them, they acknowledged several parts that we grouped into categories such as Academic Discourse, Pronunciation, among others with a preference to the categories of Organisation and Achievement. Here, we also noticed that some raters provided responses that were off-topic and others referred to dimensions/descriptors that the rubric actually lacks as shown in Table 4. This suggests again that raters do not possess enough knowledge about the rubric they are using.

Table 4

*What is the part of the oral exams that is most difficult to evaluate for you? (taking into account the rubric)*

| Categories | Frequency |
|---|---|
| Organisation | 4 |
| Achievement | 2 |
| Pronunciation | 1 |
| Interaction | 1 |
| Register | 1 |
| Academic Discourse | 1 |
| Rehearsed Discourse | 1 |
| Linguistic Range | 1 |
| Lexicon | 1 |
| Accuracy | 1 |

Regarding the type of English the raters evaluate in the oral exams, they indicated contents labeled Social English, while the Academic/Professional English appeared 3 times and Technical English was indicated by one rater. Results suggest that there is a consistency between the raters' answers in relation to the non-existent alignment and the social English evaluated in oral exams (see Table 5 below).

Table 5
*What type of English do you evaluate in tests?*

| Type of English | Appearance |
|---|---|
| Social | 12 |
| Academic/Professional | 3 |
| Tecnical | 1 |

When raters were asked if the rubric corresponded with to what raters teach/practice in the course and what they evaluated in the oral exams, almost every rater expressed their opinion that the rubric is indeed aligned with the programme and what is taught in the classroom. This indicates a positive perception of the construct validity of the rubric. However, the rater that responds negatively argued with an answer off-topic because it was not related to the rubric (see Table 6).

Table 6
*Do you think that the rubric corresponds to what you teach / practice in the course and what you evaluate in the oral exams?*

| Answers | Number of Raters |
|---|---|
| Yes | 13 |
| No | 1 |

When the raters were asked if they would add or remove any dimension to the oral exam rubric, 13 out of 14 would not add any dimension to the actual rubric (see Table 7). One of the 13 raters that previously answered that he/she would not add any

Table 7

*Do you add or remove any dimension to the oral exam rubric?*

| Answers | Frequency |
|---|---|
| Yes | 1 |
| No | 13 |

on to the rubric provide another answer but her/his answer is off-topic, not related to the question and one rater referred to a dimension that is not included in the rubric and that was categorised as 'Appropriacy' (see Table 8). This dimension seemed to be related to the objective of penalising rehearsed discourse. In this particular response, the rater probably was referring to a property closer to the idea of Authenticity. For practical purposes, this category will be referred to as Authenticity along the study.

Table 8

*Which one?*

| Dimension | Frequency |
|---|---|
| Authenticity | 1 |
| Off-topic | 1 |

Concerning oral ability, the tendency among raters' perceptions was to refer to extra-rubric dimensions such as 'Pronunciation' and 'Appropriacy' (meaning Authenticity). This actually suggests discomfort with the actual rubric or lack of knowledge about it. Additionally, there were two responses that indicated dimensions such as 'Non-verbal expressions' and 'General performance', which already existed in the actual rubric as 'Achievement' (see Table 9). This also suggests little knowledge about the rubric, its dimensions and, consequently, the constructs it describes for assessment purposes.

Table 9

*What aspects of the students' oral ability would you add to the rubric?*

| Aspects | Appearance |
|---|---|
| Do not add any dimension | 4 |
| Pronunciation | 4 |
| Appropriacy | 3 |
| Non-verbal expression | 1 |
| General Performance | 1 |

Note 5: 1 rater points out he/she would add a 'General performance' dimension, regarding the oral ability. However, he/she likely referred to the 'Achievement' dimension that already exists in the actual rubric.

All the raters proved to know the level of English expected in the course and agreed that students should reach B1 at the end of the programme, which is totally aligned with what the programme says (see Table 10 and 11).

Table 10

*Do you know what level of English is expected in the current course (s) you teach at the university?*

| Answers | Frequency |
|---|---|
| Yes | 14 |
| No | 0 |

Table 11

*Which one?*

| Levels | Frequency |
|---|---|
| A1 | 0 |
| A2 | 0 |
| B1 | 14 |
| B2 | 0 |
| C1 | 0 |
| C2 | 0 |

However, only half of the raters believed that the students actually reach that level of English (see Table 12). In this way, it is possible to conclude that, according to the raters' perception about the oral exam rubric, a lack of construct validity can be inferred. This suggests a significant inconsistency among raters when comparing these results to those regarding face validity, as discussed below (see section 4.1.3).

Table 12
*Do you think that students reach the level of English proposed by the current course programme?*

| Answers | Frequency |
|---|---|
| Yes | 7 |
| No | 7 |

Finally, when the raters were asked to evaluate the design of the oral exam rubric in a scale from 1 to 4 (1 being the lowest score and 4 the highest score), they rated the design of the oral exam rubric with 3 or the highest score as shown in Table 13. It is also interesting that the design of the rubric is well-assessed by raters, taking into account previous suggestions from raters where they pointed out to modifications they would introduce the rubric.

Table 13
*Final rating of the design of the oral exam rubric*

| Evaluation | Number of Raters |
|---|---|
| 4 | 12 |
| 3 | 2 |
| 2 | 0 |
| 1 | 0 |

Note 2:all the raters qualify positively the design of the rubric, however only one of them participated in the design of the rubric.

### 4.1.2 Results for Research Question 1.1: Raters' perception regarding face validity

R

eTable 14

g*Are you satisfied about the oral exam rubric?*

| a | Answers | Frequency |
|---|---|---|
| r | Yes | 13 |
| d | No | 1 |
| i | | |

n

g the final evaluation of the rubric, the face validity of the rubric concerning raters' perceptions is satisfactory. It is important to notice here that raters do not seem completely satisfied with the instrument when asked about its alignment with the programme, but they still assess the rubric well when face validity is observed (see Table 14).

Table 15

*Have you participated in the design of the oral exam rubric?*

| Answers | Number of Raters |
|---|---|
| Yes | 1 |
| No | 13 |

At the same time, it is interesting to notice that only one rater participated in the design of the actual rubric (see Table 15). This also suggests problems with the face validity of the oral exam rubric, because it partially explains the lack of knowledge of the actual rubric by raters. This feature of the data set has also implications for the analysis of the rubric that is presented later in section 5.4.1. When he/she was asked about what part of the design of the rubric was more difficult at the moment of its elaboration, he/she pointed out that Generalised descriptors and Dimensions were more complex to design at the moment of the rubric elaboration.

According to the results in Table 16, most of the raters believed that the programme reaches the proposed goal even though they do not believe that students are reaching the proposed level of English by the programme (see Table 12). In line with the results indicated in Table 21, this reinforces the idea that raters have poor knowledge about the programme and/or the rubric.

Table 16
*Do you think the programme reaches the proposed goal?*

| Answers | Frequency |
|---|---|
| Yes | 11 |
| No | 3 |

There is a total agreement among raters when it comes to the usefulness of the educational tools that the programme provides to the students and about the utility of it. The main reasons provided by raters were that the educational tools were complementary among them and were very helpful since the hour classes the programme has were not considered enough to teach the content and abilities they need to. This confirms that the educational tools of the programme have positive face validity, in that raters' perceptions revolve around the improvement of different abilities (linguistic, self-study and communicative abilities) (see Table 17).

Table 17
*Do you think that the educational tools provided by the programme are useful?*

| | Answers | Frequency |
|---|---|---|
| R | Yes | 14 |
| | No | 0 |

e

sponses also showed that the raters believed that the educational tools the programme provides to the students were useful. Their explanations could be grouped into ideas such as Improvement of linguistic abilities, Improvement of self-study abilities and Improvement of communicative abilities. There were 3 raters that did not answer this question. Along the same line, the main reasons raters provided to argue for the usefulness of the educational tools were that such tools served the purpose of Complementary tools, Didactic tools, and Self-study tools. Only one rater provided an answer that was off-topic and one rater did not answer this question.

When raters were asked to grade the effectiveness of the programme where the efficiency level goes from lowest (1) to highest (4), most of them graded the programme positively, 3 was the lowest grade 6 raters provided as it is depicted in Table 18. Overall, the face validity of the rubric and the programme is satisfactory. Although this should be interpreted as enhancing the validity of the rubric, in the case of this study, it must be interpreted as revealing inconsistencies between the raters' perceptions of construct validity and their perceptions regarding face validity.

Table 18
*Effectiveness of the programme*

| Level of efficiency | Frequency |
|---|---|
| 1 | 0 |
| 2 | 0 |
| 3 | 6 |
| 4 | 8 |

Note 8: the efficiency level goes from lowest to highest, being 1 the lowest and 4 the highest.

### 4.1.3 Results for Research Question 1.1: Raters' perception regarding reliability

Results also show that raters' experience as raters goes from 2 to 6 years, which indicates that all raters have at least some experience using the rubric as an assessment instrument (see Table 19).

Table 19
*How many years do you have of experience as a rater?*

| Years of experience | Number of Raters |
|---|---|
| 6 | 2 |
| 5 | 4 |
| 4 | 2 |
| 3 | 3 |
| 2 | 3 |

In relation to the way in which the raters applied the rubric, most of them believed that the rubric was applied consistently between raters (see Table 20). This belief is supported by reasons like the well-designed rubric, the standard character of the oral exam and the rater's experience regarding rubrics. In turn, this view can be interpreted as imprecise, considering the indications that construct validity may not be good in the rubric (as seen in section 4.1.2, below).

In contrast, the raters that did not believe the rubric was applied consistently among raters based their answers on the different criteria and subjectivity of every rater. As discussed in 2.3.2, in order to have a reliable assessment, raters have to be consistent with themselves (intra-rater reliability) and among them (inter-rater reliability). Therefore, these explanations by raters regarding the application of subjectivity and individual criteria of raters may be indicating a negative influence on the reliability of the rubric.

Table 20

*Do you think every rater apply in the same way the rubric of the English programme?*

| Answers | Number of Raters |
|---|---|
| Yes | 10 |
| No | 4 |

In regard to the knowledge about the existence of the rubric, several issues became apparent. Most of the raters make the existence of the rubric known to the students, which contributes to a high degree of reliability. However, 8 raters indicated that they provided access to the rubric before the evaluation and the rest accepted that they did not do it (see Table 21 and 22).

Table 21

*Do you inform about the existence of the rubric of the oral exam to the students?*

| Answers | Number of Raters |
|---|---|
| Yes | 13 |
| No | 1 |

Table 22

*Do students have access to the rubric before the evaluation?*

| Answers | Number of Raters |
|---|---|
| Yes | 8 |
| No | 6 |

Nevertheless, 13 raters reported that they explained the rubric to the students before the evaluation and only one accepted that he/she did not do it. (see Table 23). These responses show a lack of consistency amongst the raters surveyed, since in

the previous question (see Table 22) 8 out of 14 had admitted that they did not provide access to the rubric to the students before the evaluation and, consequently, they could not have explained the rubric to the students before the evaluation. This affects directly the degree of reliability of the assessment because the preparation that the test needs is not the same for all the students. Nonetheless, all the raters assured that the rubric was applied in every oral exam, which contributes positively to the degree of reliability (see Table 24).

Table 23
*Do you explain to the students the rubric of the oral exam?*

| Answers | Number of Raters |
|---|---|
| Yes | 13 |
| No | 1 |

Table 24
*Are there oral exams where the rubric is not used to evaluate?*

| Answers | Number of Raters |
|---|---|
| Yes | 0 |
| No | 14 |

Concerning raters' training, when the raters were asked "How do you train to use the rubric consistently?", the answers showed that only one rater was formally trained in the use of rubrics. However, this training was not institutional nor was it focused on the actual rubric analysed, as we can see in his/her answer "Tuvimos un workshop una vez y evaluamos a unos estudiantes de unos videos y despues cotejamos nuestra evaluación con la que Cambridge hizo. Lo demás es simplemente práctica, ya que tenemos dos exámenes orales: uno de medio semestre y otro final".

The rest of the raters admitted that they self-trained or did not train at all. They reported justifications such as “La reviso muchas veces”, or “No se entrena”. In addition, 3 raters provided answers that were off-topic and 3 raters did not answer the question (see Table 25). These results indicate a serious threat to the reliability in the application of the rubric because of the lack of procedures for training that are required to ensure reliability discussed in section 2.3.3. This lack of training opens a clear possibility that the rubric is used according to the individual raters’ discretion, which is the opposite of what the purpose of a rubric is.

Table 25
*How do you train to use the rubric consistently?*

| Answers | Frequency |
|---|---|
| Self-training | 4 |
| Non-training | 3 |
| Off-topic | 3 |
| Rater training | 1 |

Note 3: Not all the raters answered the question.

Finally, raters reported a different variety of comments about the most important dimensions for them, such as ‘Linguistic Range’, ‘Accuracy’ and ‘Achievement’. It is significant that the dimensions to which the raters gave more importance were Linguistic Range and Accuracy, as these are the same dimensions that were observed to be overlap (see section 4.3 below. Responses here suggest again that raters do not use the rubric consistently, because they consider some dimensions more important than others. At the same time, there were two raters that provided answers that referred to dimensions not present in the rubric such as Orthographic aspect and English-Grammar use (see Table 26). This indicates again lack of knowledge of the rubric from the raterd Only three raters reported that they focus on all the dimensions equally, which generates further expectations of unreliability in the use of the rubric.

Table 26
*What aspects of the rubric do you consider most important when evaluating?*

| Aspects | Appearance |
|---|---|
| Linguistic Range | 4 |
| Accuracy | 4 |
| Achievement | 4 |
| Organization | 3 |
| All dimensions | 3 |
| Orthographic Aspect | 1 |
| English-Grammar use | 1 |

Note 4: the raters provide more than one aspect in their answers.

## 4.2 Results for Research Question 1.2: Students' perceptions about the rubric

Under RQ1.2, three sub-questions were posed to observe the perceptions of the students regarding three different aspects of validity: construct validity, face validity, and reliability in relation to the oral exam rubric. Complete data sets are reported in Appendix B.

### 4.2.1 Results for Research Question 1.2: Students' perception regarding construct validity

According to the student's answers regarding if they knew the English level expected in their current course, 13 replied they knew while 10 did not. Out of those who knew, 9 of them indicated that they are in the English IV level, 3 indicated that they are at an Intermediate level and 1 indicated that they were at an Advanced level. Interestingly, nearly half of the students failed to acknowledge their expected English

level, which suggests a lack of awareness the students have in regards to their English programme and level.

Along the same line, 11 out of 23 students claimed they did not believe they would achieve their expected English level in their current course. This suggests either a lack of confidence the students have towards their respective English programme and its effectiveness, or a lack of trust in their own abilities. The rest of this section's results seem to point out to the former option. Next, only 11 out of the 23 surveyed students considered that their course programme adapted to the needs of their respective undergraduate programmes. Although the Blended Learning English programme is an across the curriculum type, it is not perceived by students as accommodating adequately to their particular undergraduate programmes.

As seen in Table 27, when asked what part of the oral exams they thought of as the most difficult for them was, the students' answers were varied. However, the majority recalled Grammar as the most difficult part of their evaluations. The next most difficult item for them was Vocabulary with 5 answers, followed by Fluency with 4. Other difficult items reported by the students were Pronunciation with 3 answers; Register, Discourse, and Relevance of information with 2 answers; and finally Authenticity with 1 answer. What called our attention from these answers is the overall unrelatedness from the reported items to what the rubric is supposed to evaluate. That is, items like Grammar, Pronunciation, and Fluency are not present in the oral exam rubric. This could be interpreted as evidence of weak programme-rubric alignment.

Table 27

*What is the part of the oral exams that is most difficult for you?*

| Categories | Frequency |
|---|---|
| Vocabulary | 5 |
| Grammar | 9 |
| Register | 2 |
| Discourse | 2 |
| Pronunciation | 3 |
| Fluency | 4 |
| Authenticity | 1 |
| Relevance of information | 2 |

Next, as depicted in Table 28, when the students were asked what type/s of English is/are evaluated in the oral exam of the course, nearly all of them replied as being Social, which means it is used with the aim to express themselves in a colloquial environment, such as a friendly conversation or asking for directions. Other students also listed the English taught in the programme as Academic/Professional, and others as being Technical. The former refers to academic discourse that requires a formal use of language, for example, the writing of a paper or the elaboration of a congress presentation, while the latter refers to the specific vocabulary in a particular field or area of knowledge, for example, to give a diagnosis if one is in medical school. This can be interpreted as the programme focusing its attention towards the minimum skills the students require in order for them to develop basic communicative skills in English, while at the same time passing over specific contents (technical vocabulary, for example) related to their particular undergraduate programmes.

Table 28
*What type/s of English is/are evaluated in the oral exam of the course?*

| Categories | Frequency |
|---|---|
| Académico / Profesional | 5 |
| Social | 22 |
| Técnico | 6 |

Further on, the majority of the students agree that Communicative skills convey the prime example of the type of English that is evaluated in their oral exams, as in a conversation with a patient or to give and ask for directions. 6 students reported Personal affairs, such as describing family holidays or their daily routines, and lastly 4 students responded referring to Authenticity, for example referring to the fluidity in their speech. Although many students reported on the same item (even though each reported it differently), this lack of consensus further suggests the idea that they were not informed enough regarding their course programs and what it is expected from them, and so the construct validity of the rubric is compromised.

Table 29
*Can you give examples of the type/s of English evaluated in your oral exam?*

| Categories | Frequency |
|---|---|
| Communicative skills | 15 |
| Personal affairs | 6 |
| Authenticity | 4 |
| Descriptions | 1 |
| Spelling | 1 |
| Off topic | 1 |

Note 10: There are stakeholders that mentioned more than one example, which is why the number of answers is more than 23.

We then proceeded to ask the students if they believed the oral exam rubric describes what is taught/practiced in the course. Interestingly enough, the majority of students think the rubric is a good representative of what they are taught in their courses. Nonetheless, these answers do not correlate with previous questions they were asked, since, so far, the students had reflected mostly unawareness regarding both their programmes and its oral exams rubric. These results may suggest that students are either compromising their answers in order to avoid the responsibility to back them up or that they are in fact sure about the effectiveness of the rubric, which does not seem to be the case (see Table 30).

Table 30
*Do you think that the oral exam rubric describes what is taught / practiced in the course?*

| Answers | Frequency |
|---|---|
| Yes | 17 |
| No | 6 |

Along the same lines, 16 students thought that the rubric corresponded to what was evaluated in the oral exams, while only 2 disagreed and 5 students were undecided. This seems to corroborate that the majority of students agreed that the programme is well represented in the rubric dimensions, and also that the rubric is a good assessment tool for their oral evaluations in terms of correspondence. This, however, seems inconsistent on the part of the students since they seemed to evaluate very positively the rubric used for their oral evaluations while at the same time demonstrating they were unaware of their programme and what should be evaluated by the rubric in their oral exams (see Table 31).

Table 31
*Do you think that the rubric corresponds to what is evaluated in the oral exams?*

| Answer | Frequency |
|---|---|
| Yes | 16 |
| No | 2 |
| Undecided | 5 |

Finally, as seen in Table 32, the students were asked if they would change something from the rubric in order to check their levels of affinity towards it, that is whether they were acquainted with the rubric for their oral exams or not. The results, as pictured in Table 43, show that only 6 people said yes, 4 were undecided, and 13 simply said no. These answers are in accordance with the previous overall satisfaction towards such rubric. Although most students would not change anything from the rubric (meaning any change would be deemed as unnecessary and thus affirming the effectiveness of the rubric), we can still interpret these answers as not explanatory enough.

As previously reported in this section, most students lack awareness regarding both their programmes and the rubric for oral exams. This poses the question of why students would think so highly of their rubric for oral exams while knowing relatively little about their programmes and thus about the rubric itself. We will try to unravel this issue further on in section 4.2.2.

Table 32
*Would you change something from the oral exam rubric? What?*

| Answers | Frequency |
|---|---|
| Yes | 6 |
| No | 13 |
| Undecided | 4 |

### 4.2.2 Results for Research Question 1.2: Students' perception regarding face validity

In regard to the student's face validity perceptions of the rubric, almost all students indicated they felt that there were aspects of their oral skills that were not being evaluated in the rubric (see Table 33). Only one student said yes and referred to Authenticity and Relevance of Information as an explanation (See Table 34).

Table 33
*Do you think there are important aspects regarding your oral skills that are not evaluated in the rubric?*

| Answers | Frequency |
|---|---|
| Yes | 1 |
| No | 22 |

Table 34
*Which?*

| Categories | Frequency |
|---|---|
| Authenticity | 1 |
| Relevance of Information | 1 |

Note 13: 1 of the answers corresponds to a *yes* answer from table 47 and the other corresponds to a *no* answers from the same table.

This suggests that the majority of the students believe that the rubric actually measures everything it has to measure in relation to oral skills. This indicates that the rubric has good face validity, understood as the degree to which a test looks right and appears to measure the ability that is intended to measure appears to measure (as

explained in section 2.2.1). Nevertheless, these results must be taken cautiously as 12 out of 23 students claimed also not to have access to the oral exam rubric before an instance of evaluation (see Table 40). This means that they were not able to see what the rubric actually looks like and what is actually measuring and, therefore, they could not know with certainty if there are not aspects regarding oral skills missing in the evaluation that they are receiving.

Further on, in Table 35, 17 students indicated that they considered that the oral exam rubric measures effectively the oral skills, while 6 students indicated that it does not (see Table 35).

Table 35
*Do you consider that the oral exam rubric measures effectively the oral skills?*

| Answers | Frequency |
|---|---|
| Yes | 17 |
| No | 6 |

These results indicate again a high degree of face validity of the rubric. However, considering the issues regarding construct validity that the rubric presented (see section 4.2.1), and the lack of access they have to the rubric (see section 4.2 3), it is clear that the students were not aware of the inconsistencies and ambiguities that the rubric actually has. By any means, as pointed in the literature, face validity depends on how the test-taker or the test-giver, the students, in this case, perceive the instrument. Therefore, even though the rubric may present issues of construct validity (see section 4.2.1), it can still have high levels of face validity as long as the students approve of it for reason different to its construct validity.

Students were also asked if the programme met their expectations, to which 13 agreed but 10 indicated that it had not (see Table 36). According to those students who answered positively, 7 thought that the programme met their expectations due to the general improvement of their oral skills. Those students whose answers were no,

attribute their answers mainly to pedagogical issues such as unsuccessful teaching strategies, inadequate elaboration of the content and logistic issues such as lack of time. This is due to the fact that the students who answered no felt that they did not manage to acquire successfully the contents and abilities that they were supposed to during the course/s.

Table 36
*Has the programme met your expectations?*

| Answers | Frequency |
|---|---|
| Yes | 13 |
| No | 10 |

It is here interesting to notice a certain contradiction between the number of students that assert that the programme did not meet their expectations and the number of positive feedback that the students gave the programme in other questions such as Do you consider that the oral exam rubric measures effectively the oral skills? or How effective do you think the across the curriculum English programme is? (see section 4.2.1). This inconsistency can be attributed to either the lack of knowledge that the students had in relation to the programme and therefore, what to expect from it; or the lack of willingness from the students at the moment of answering seriously and consistently the questions.

Next, as observed in Table 37, we can infer rather high levels of face validity from the students regarding the rubric despite the many issues that were encountered in terms of the rubric's construct validity. Out of 23 students, 16 indicated that they were satisfied with the rubric they use for oral evaluations, while only 7 were discontent with it (see Table 37). Students attributed their negative answers to rubric aspects such as lack of authenticity, lack of familiarization with the rubric, automatization issues and improvements that the rubric should have according to the programme in which it is being applied. Students that answered positively referred

mainly to features of the rubric such as its usefulness, its alignment with the programme and its quick and accurate feedback.

Table 37
*Are you satisfied with the oral exam rubric?*

| Answers | Frequency |
|---|---|
| Yes | 16 |
| No | 7 |

The 16 students represent the 69,6% of the total amount of answers, which can be considered as a very high percentage that firmly supports the face validity of the oral exam rubric. This means that students had a positive general appreciation of the programme. However, this general appreciation does seem coherent with the construct validity and reliability problems that were identified in this study (see 4.2.1 and 4.2.3 respectively). Therefore, even though the levels of face validity of the rubric are very positive, this may not be a reflection of the actual quality of the rubric.

The majority of students also agreed on the educational tools provided by the programme is useful and only 7 of them felt that they were not useful at least to some degree (see Table 38).

Table 38
*Do you think that the educational tools provided by the programme are useful?*

| Answers | Frequency |
|---|---|
| Yes | 16 |
| No | 7 |

These results demonstrate again that the students were rather comfortable with the programme and with what it has got to offer as they considered the tools provided by it to be useful and satisfactory. This boosts the face validity of the programme since, at first sight at least, it seems to provide enough amount of learning tools.

Based on 4 levels of effectiveness, a vast majority of students agreed that the programme was good. One student qualified the programme with the highest grade. Similarly, only one student rated the programme with the lowest score, and the remaining 4 students scored the effectiveness at a level 2 out of 4 (see Table 39). This suggests that students had a rather positive general appreciation of the programme which is in accordance with the high levels of face validity of both rubric and programme.

Table 39

*How effective do you think the across the curriculum English programme is?*

| Answer | Frequency |
|---|---|
| 1 | 1 |
| 2 | 4 |
| 3 | 17 |
| 4 | 1 |

Along these lines, these results demonstrate that construct validity is not necessarily related to face validity, that they are independent constructs and thus illustrates the multifaceted nature of validity, as put by Messick (1989). Consequently, it can be said that the face validity of the oral exam rubric and the across the curriculum English programme concerning students' perceptions is acceptable and satisfactory. Nonetheless, it must also be noticed that in other questions of the survey, the students do not seem to be completely satisfied with the programme or the rubric and they also show a lack of knowledge of both.

### 4.2.3 Results for Research Questions 1.2: Students' perception regarding reliability

11 Students indicated that they had access to the rubric before an evaluation, while 12 students, which is more than the half, did not have access to the rubric before an evaluation (see Table 40).

Table 40
*Do you have access to the rubric before an evaluation?*

| Answers | Frequency |
|---|---|
| Yes | 11 |
| No | 12 |

A

lso, 20 students indicated that they had been explained the rubric by the raters while only 3 of them stated that the rubric has not been explained by their teachers. This shows that there is a clear inconsistency when students and raters are asked about the awareness of the rubric (see Table 41).

Table 41
*Have you ever been explained any rubric of the oral exam?*

| Answers | Frequency |
|---|---|
| Yes | 20 |
| No | 3 |

Students were also asked if they knew any oral exam in which raters had not used the rubric. 18 of them answered that they do not know of any instance and 5 of them answered they did (see Table 42). The students were asked if the raters use the

rubric in the same way. 14 of them stated that raters do use the rubric in the same way while 9 of them stated that raters do not use the rubric in the same way (see Table 43). This suggests that raters use the rubric differently, which is coherent with lack of training for the implementation and constitutes a threat to the validity of the test.

Table 42

*Do you know instances of oral exam in which your teachers have not used the rubric?*

| Answers | Frequency |
|---|---|
| Yes | 5 |
| No | 18 |

Table 43

*Do all your raters use the rubric in the same way?*

| Answers | Frequency |
|---|---|
| Yes | 14 |
| No | 9 |

Finally, 17 students thought that raters prioritize certain aspects over others at the moment of evaluating, while 6 students thought raters did not prioritize aspects (see Table 44). These results suggest that raters present low levels of reliability among each other since they tend to interpret the rubric in their own individual ways. In addition to this, these answers indicate that the rubric is unclear as raters may get confused and interpret the dimensions differently affecting the validity of the test.

Table 44

*Do you think that the raters prioritize certain aspects over others at the moment of evaluating?*

| Answers | Frequency |
|---|---|
| Yes | 17 |
| No | 6 |

## 4.3 Results for Research Question 1.3: Rubric features regarding construct validity

The USACH rubric is constituted by five dimensions that indicate the learning objectives of the assignment. The dimensions identified in the rubric are 'Achievement', 'Linguistic range', 'Organization', 'Accuracy', 'Register and Academic Discourse'. The description of the dimensions provides an explanation of what is anticipated for each dimension on the scale. The levels/bands are six in total, and they describe the qualities required to demonstrate achievement of each standard for each criterion, from highest to lowest.

In order to answer RQ3, regarding rubric features about construct validity, the main findings are presented in relation to the components of the rubric, i.e: dimensions, levels/bands, descriptors, and alignment.

### 4.3.1 Dimensions

Regarding the dimensions of the rubric for oral exams, there are confusing levels of ambiguity in the description of the dimensions that may imply bias at the moment of evaluating and grading. For instance, we considered that the dimensions 'Linguistic range' and 'Accuracy' are described in a similar way, producing an overlap of the content of the dimensions. Accordingly, there are

dimension overlap problems related to the construct validity of the rubric that might affect students' performance and also teachers' grading.

Table 45
*Dimensions comparison*

| **Accuracy** | **Linguistic Range** |
|---|---|
| **Did the learner produce grammatically correct language?** | **Did the learner use a wide variety of vocabulary and grammar structures?** |
| No errors in use of structures and vocabulary expected at this level | A wide variety of both correct structures and appropriate vocabulary used |
| Very few errors in use of structures and vocabulary expected at this level | A wide variety of both structures and vocabulary used appropriately with minor difficulty |
| **Some elements of 4 and some of 6** | **Some elements of 4 and some of 6** |
| Errors in use of structures and vocabulary are common, but rarely impair communication | A variety of appropriate structures used, with some inappropriate use of language |
| Some elements of 2 and some of 4 | Some elements of 2 and some of 4 |
| Frequent errors make learner's writing difficult to understand | Vocabulary and structures used are very limited in range and often inappropriate |

As shown in Table 45 above, both dimensions state something very similar. For instance, they establish that the learner has to produce grammatical structures correctly. Under Accuracy, the students cannot make errors in vocabulary and grammar structures, which is similar as in the linguistic range dimension, where the students are evaluated according to the variety and appropriate structures and vocabulary features. Both dimensions, therefore, penalize errors in the same features

of language (grammar and vocabulary). To overcome this overlap, these dimensions should be treated as one integrated dimension and not separately or described in ways in which the two constructs are clearly distinguished from each other

### 4.3.2 Levels or Bands

We also observed that 2 levels relied on the content of the descriptor of the levels of their neighboring levels. The descriptions indicate in these levels the combination as "some elements of 4 and some of 6". This means that these levels are not actually described as their description has to be borrowed from other descriptors.

Table 46
*Levels or Bands*

| | | | |
|---|---|---|---|
| A wide variety of both correct structures and appropriate vocabulary used | Learner able to connect ideas clearly and effectively, using linkers and devices appropriate to the level | No errors in use of structures and vocabulary expected at this level | The learner is able to use appropriate language when presenting at different levels of formality. |
| A wide variety of both structures and vocabulary used appropriately with minor difficulty | Learner able to connect ideas clearly and effectively, generally using linkers and devices appropriate to the level | Very few errors in use of structures and vocabulary expected at this level | The learner is able to use appropriate language when presenting at limited levels of formality. |
| **Some elements of 4 and some of 6** | **Some elements of 4 and some of 6** | **Some elements of 4 and some of 6** | **Some elements of 4 and some of 6** |
| A variety of appropriate structures used, with some inappropriate use of language | Learner usually able to communicate and link ideas clearly, though sometimes errors make meaning unclear | Errors in use of structures and vocabulary are common, but rarely impair communication | Meaning is largely clear, but the learner is not able to choose language according to register |
| **Some elements of 2 and some of 4** | **Some elements of 2 and some of 4** | **Some elements of 2 and some of 4** | **Some elements of 2 and some of 4** |
| Vocabulary and structures used are very limited in range and often inappropriate | Learner only able to use very basic linking devices, and meaning is often unclear | Frequent errors make learner's writing difficult to understand | Too little communication to assess |

The problem with these two levels is that because they rely on the descriptors of their neighbouring levels, which are sometimes poorly described (see section 4.3.3 below), the appreciation of the performance level becomes a function of the capacity of the rater to combine descriptions belonging to two different levels. It can be safely be argued that such levels actually do not exist, as there is no description of them. This poses an important problem for the validity and reliability of the rubric since it does not help in observing students' performance and promotes individual interpretations of raters.

### 4.3.3 Descriptors

Another important issue is related to the descriptor of the Linguistic range dimension, which is stated as follows: "Did the learner use a wide variety of vocabulary and grammar structures?", where the word "wide variety" is not clear enough when it comes to knowing the needed, suggested or requested number of specific words or structures that the students have to produce in order to accomplish the expectation of the dimension successfully.

Table 47
*Linguistic Range*

| **Linguistic range** |
| --- |
| **Did the learner use a wide variety of vocabulary and grammar structures?** |
| A **wide variety** of both correct structures and appropriate vocabulary used |
| A **wide variety** of both structures and vocabulary used appropriately with minor difficulty |
| Some elements of 4 and some of 6 |
| A **variety** of appropriate structures used, with some inappropriate use of language |
| Some elements of 2 and some of 4 |
| Vocabulary and structures used are very limited in range and often inappropriate |

Table 47 shows that the rubric does not present a progression or scale in the degree of performance (from highest to lowest) in the descriptors used in relation to the quality of variety. The rubric shows only "wide variety" and then "variety", which are not easy to interpret for assessment purposes. This is also an issue in the evaluation mainly because these descriptors do not seem to be fit for students in the English I or II courses as observed in section 4.3.4. (See appendix D)

### 4.3.4 Alignment between rubric and course programme contents

The rubric under analysis is used to evaluate the four courses of the across the curriculum programme and thus there is not alignment between the descriptions of the rubric and the contents that are taught on the courses English I and English II. For instance, in the dimension Linguistic Range, to achieve the maximum

score it is required the use of a "wide variety of correct structures". However, for the English I course, the student is expected to learn two tenses, Past Simple and Present Simple. Then, in the English II course, the student is introduced to other two tenses, Past Progressive and Future (will and going to). The requirements of this rubric are therefore not adapted to the contents of each programme.

According to the planification of the courses and the descriptors of the rubric, they are aligned to more advanced courses, such as English III and IV, where the students manage more tenses, such as narrative tenses and complex structures (active and passive voice).

For these reasons, it is possible to state that this rubric looks in this sense more like a scoring guide rubric rather than a rubric that evaluates the progress of the student in each course. This rubric seems more like a tool that evaluates the performance of the task in the oral exam, which in turn does not seem to include elements of the programme itself. This issue can be observed in the description of the dimensions of the rubric because it is nod adapted to the contents of the different levels of the programme.

The evidence suggests here that the rubric for oral exams presents a low degree of construct validity, since all its components present some issues in the description of the abilities measured in the rubric. This can be illustrated with the analysis of the levels, descriptors, and alignment reported above, which shows critical problems on all dimensions of the rubric, and also with the perceptions of both, the students and raters.

In summary, considering all these observations, it could be expected that stakeholders and raters could detect some of these issues that affect construct validity directly, either in the ambiguity and imprecision of the concepts, or any other related to issues of application. As seen in 4.3, this was not the case, as the face validity of the rubric was quite positive.

## 4.4 Similarities and/or differences between raters' and students' perception about the oral exam rubric

For practical purposes, the surveys were elaborated with mirror questions that allowed us to observe the similarities or differences that may exist between raters' and students' perceptions.

Results observed for RQ1.1 (section 4.1) and RQ1.2 (section 4.2) show that there is an agreement between students and raters with respect to the type of English evaluated in the programme. The majority of the stakeholders points out the type of English is Social, which supports the idea that raters are teaching the same content the students perceived that they are learning in the programme. This suggests a favorable degree of construct validity in this case.

Following with the analysis, in relation to the expected level of English of the programme, contrasting results between raters and students can also be observed. On the one hand, all raters reported a complete knowledge about the expected level of English proposed by the programme; on the other hand, 10 out of 23 claimed not knowing the expected level of English proposed by the programme.

This may be indicating that there are problems with the information provided to students regarding the programme and the assessment instrument. This may imply, in turn, that raters are not being clear about this issue with the students. Most raters believe that the programme is not adapted to the needs of every programme, otherwise, students perceptions regarding this issue are divided. This supports the belief that this is an across the curriculum programme and do not aim to fulfill the specific needs of every programme.

Concerning the reliability of the rubric, raters and students were surveyed about their difficulties at the moment of evaluating and being evaluated respectively. The responses to these questions were quite varied. However, it is worth noting that some answers pointed to dimensions that are not present in the rubric, such as 'Pronunciation', 'Interaction', 'Rehearsed discourse' and 'Lexico', in the case of

raters. On the other hand, students mentioned aspects such as ‘Pronunciation’, ‘Authenticity’ and ‘Relevance of information’, which are not dimensions of the rubric either. This may suggest lack of knowledge of the assessment instrument by raters and students and, consequently, problems with the reliability of the rubric. These problems arise from the fact that raters are assessing oral exam performance without having an optimal knowledge of the rubric and students are being evaluated with varying degrees of information regarding the rubric.

Along the same line, most of the raters answered that they actually applied the rubric in the same way; however, 4 raters responded negatively to this question. Students, in turn, were divided in their answers: 14 out 23 reported positively to this question and 9 negatively. This may indicate a low degree of reliability in relation to the rubric, due to the lack of consistency between raters, perceived by raters and students.

There is a total agreement between raters and students that the rubric corresponds to what is taught and evaluated in the course syllabus. This indicates a satisfactory face validity that raters and students have of the instrument and the programme. However, such high face validity is inconsistent with the issues within the construct validity of the rubric observed in sections 5.1.1 and 5.2.1.

It is interesting that when it comes to answer whether students have access to the rubric before an evaluation, responses of raters and students show varied practices. On the one hand, 8 out of 14 raters claimed that they provided access to the rubric to students before an evaluation. On the other hand, 11 out of 23 students indicated that they did have access to the rubric before an evaluation. These results indicate that there is a lack of consistency among the raters surveyed which is also reflected in what students reported. Raters that are not providing adequate access to and explanations of the rubric to the students affect the validity of their assessment as the evaluation instrument decreases in its reliability.

Similarly, 20 out of 23 students answered they have been explained the rubric for the oral exam by the raters and 13 out of 14 raters report to have explained the

rubric to students. In spite of the fact that responses by raters and students here are consistent, as mentioned in Section 5.1.3 and 5.2.3, students' answers regarding the awareness of the rubric are not so. As shown in Table 40, 12 out of the 23 students did not have access to the rubric before an evaluation. This may imply that even though raters answered they all do explain the rubric to students; some raters do not ensure that students get acquainted with the instrument. This is a problem of reliability that directly affects the validity of the assessment procedures, as some students are not properly trained for the oral exam.

Next, in sections 4.2.3 and 4.1.4, it can be noticed that there is an inconsistency between raters' and students' answers. None of the raters responded that they have evaluated students' oral exam without using the rubric, but 5 out of the 23 students point out that they know instances in which raters have not used the rubric in oral exams. Furthermore, students and raters were asked whether they would change something from the oral exam rubric or not. Only 1 out of 14 raters answered that he/she would change something from the oral exam rubric and that category was more focused on the idea of Authenticity. On the other hand, most students would not change any dimension of the oral exam, yet 4 were undecided and 6 students would change something. This point illustrates that students are more likely to identify weaknesses in the rubric. When students were asked about what dimension they would add to the rubric, the most frequent feature they would add to the rubric was related to the idea Mistake penalization (students indicate that they wanted not to be penalized by making mistakes. In this case, it seems clear that the responses are inconsistent between raters and students. We believe that raters' responses indicate that the rubric has high levels of face validity as they rely on the design of the rubric but students still feel the rubric needs improvement.

According to the results observed in section 4.2.3, it can be seen that the majority of the students surveyed considered that raters give more importance to some aspects over others at the moment of evaluating, being only 6 the ones who do not consider that they did. In section 4.1.4, likewise, the results show that all raters indeed gave priority to certain aspects over others when evaluating, such as

Linguistic range, Accuracy or Achievement. These results imply consistency between what the students think about how raters use the rubric and how the raters actually proceed . The situation is thus that raters seemingly give more value to some aspects of the ability their evaluating according to their individual preferences. This affects negatively the reliability of the rubric, as its design assumes that all dimensions and aspects of language ability it evaluates are equally important in the evaluation process.

In relation to results reported in sections 4.1.1. and 4.2.1, it is possible to find a noticeable inconsistency among the answers provided by both stakeholders. In section 4.1.1, Results showed that more than half of the raters add another dimension to the rubric, such as Pronunciation or Appropriacy (meaning Authenticity), as they do not consider that the rubric measures everything it has to measure. In contrast, as shown in section 4.2.1, the vast majority of students considered that there are not aspects or dimensions being left out from the instrument. This demonstrates a noticeable difference between what raters and students think of the functioning of the rubric which, in turn, suggests that both stakeholders do not have the same level of awareness in relation to the instrument. Also, the results reported in the same sections suggest that there are problems in relation to the construct validity of the rubric, as raters believe that there are in fact aspects and dimensions missing from the instrument.

It can be stated that both stakeholders are consistent in their answers to the question: Are you satisfied with the oral exam rubric? More than half of the students indicated that they were indeed satisfied while almost all raters pointed out that they were satisfied with it. These results reflect good face validity of the test as they indicate that stakeholders consider the instrument to be good enough for its purposes. However, different answers might have been also expected considering that stakeholders indicate weaknesses of the instruments in the answers to other questions related to construct validity.

Considering how the USACH's Blended Learning programme is devised, mainly via online and with several educational tools at students' disposal, it is only correct to inquire both groups of stakeholders about their perceptions of such tools, whether they find them useful or not so much. On the one hand, the raters provided completely positive responses towards the usefulness of the programme's tools. On the other hand, the students did not seem to completely agree with that positive view of the way Blended Learning is assembled as a programme. The majority agreed that the programme tools are useful and 4 students gave off topic argumentations for their answer, which reinforces the idea that the rubric is not well-known or understood by all students. Nevertheless, some of the positive responses are justified in terms of the educative tools to provide them with new learning techniques, or implementing successful pedagogical strategies, for example. These answers seem coherent with those by the raters' who reported on the improvement of communicative, linguistic, and self-study abilities as well.

Subsequently, we decided to inquire the stakeholders for their opinions regarding the effectiveness of their across the curriculum English programme. In this item, both students and raters were asked to value the effectiveness of the programme on a scale of 1 to 4. The majority of students gave it a positive score. Raters were slightly more positive in their answers, as since 9 evaluated the programme positively.

It is interesting to see such high ratings of approval for the programme, in the face of issues observed regarding syllabus and rubrics, within the rubric, and towards the overall validity and reliability of the raters. Nonetheless, it could be argued that such appraisal towards the programme and its components may be the result of high percentages of course approval from the students, or such appraisal could simply be aided due to the unawareness the students have in relation to the problems and inconsistencies we have found so far.

Finally, the stakeholders were asked to report on whether they believed they (or their students) would achieve the expected level proposed by their current English

course programme. Surprisingly, in this item, the previous positivity towards the programme seem to slowly fade since both groups of stakeholders showed disagreement within their answers. Half of the raters believed their students did not reach the English level proposed by their programmes. Almost half of the students thought they would achieve such level, which are relatively low expectations for the success of the course.

In the next section, the main findings will be discussed regarding the issues previously addressed, in order to provide a general picture of the possible implications related to this investigation.

## 4.5 Discussion of the results

For practical purposes, the discussion of the results is organised under the three aspects that have been observed in this study: construct validity, face validity, and reliability.

The evidence provided in this study indicates that the rubric has low levels of validity. Construct validity and reliability, in particular, present several issues that contribute to low general validity although face validity is very high. This situation reflects the idea by Messick (1989) that validity is simultaneously a unitary as well as a multifaceted phenomenon (as explained in 2.2.1). This means that validity is a property of assessment that is constituted by many factors that are independent from each other, and so the validity of the rubric can be constituted by high levels of face validity, but low levels of construct validity and reliability.

From the group of raters that participated in this study, only one of them was involved in the design of this rubric and described difficulties at the moment of the elaboration, specifically with the bands and general criteria. The fact that most of the raters did not participate in the elaboration of the rubric and the fact that they do not have an instance for training or learning how to use it can reasonably explain that they did not find any issues regarding the rubric. In turn, this lack of understanding of the rubric may explain the high levels of face validity that this instrument has.

The rubric thus displayed low levels of reliability due to the fact that it does not meet the conditions of reliability established by Luoma (2004), such as standard setting and rater training procedures. Also, the fact that the raters have no training for the rubric implementation means that the rubric is not presented equally to students.

For the raters, the rubric is a useful tool to evaluate the oral exam, regardless of the problems observed, such as the overlap of dimensions and the ambiguity of the vocabulary used in the descriptors (see section 4.3). However, among students, there is a perception that some raters paid more attention to some dimensions over others. This can be explained by a lack of alignment between the programme levels and the rubric. The competencies that are required by the rubric are only achieved in English III and IV from the across curriculum programme. It seems here that in the process of adaptation of the original rubric from Cambridge, the expected progress of students' ability along the four levels of the programme was not considered adequately. This situation can be expected, as a transferring and adapting a rubric are not effective methods for rubric design, due to the fact that the rubric must be created for the particular population that is going to be evaluated (as explained in 2.2.3).

A few students reported in the survey that there are some instances where the rubric is not used in some oral evaluations (see section 4.2.3). This is inconsistent with rater's responses, who stated that there are not instances where a rubric is not used (see 4.1.4). In addition, some raters also mentioned that they only explain the rubric rather than show it (see 4.1.4). This means that the lack of knowledge of this instrument is due to the fact that students do not have access to the rubric before the test or an oral exam.

All the issues aforementioned compromise the validity of this rubric since this instrument is not aligned with the needs of the population of students nor the expectations of all courses of the across the curriculum programme or the careers that are offered in the university. This is a situation that is not described in the literature reviewed for this study, which confirms the need to expand the contexts in which the

use of rubrics are observed beyond the educational system in the United States. US experience is, clearly, not representative of all the educational systems.

Furthermore, it can be perceived that there is a general lack of awareness of the rubric that affects the validity of the test. The fact that students do not have a deep understanding of the rubric and that the raters use the rubric differently means that there is a need for procedures for ensuring reliability as rubrics need to have acceptable levels of reliability in order to achieve a good level of validity. With more knowledge about the rubric, face validity should not only improve but also would be more meaningful for rubric design.

In the same way, as raters use the rubric according to their own individual criteria and subjectivity, student's results are inconsistent. If we take into consideration that, this programme of English is meant to be standardized, a high degree of reliability is critical. In relation to the rubric, there are levels that have no meaning and dimension overlap was detected (see 4.3). This issue with the rubric allows raters to interpret the descriptors differently while assessing different students.

Regarding the face validity of the rubric, it was indeed unexpected that the rubric demonstrated a great level of face validity, considering the problems in regards to the construct validity of the instrument observed. In this sense, the features of the rubric regarding reliability and construct validity seem to be in agreement, while the face validity of the rubric seems to be disassociated with the other two aspects of validity, as it was mentioned throughout this section.

### 4.5.1 Towards a rubrics validation model

According to our review of the literature, it was observed that there is no rubrics validation model in higher education (as explained in 2.2.3). However, this study provides useful considerations for the elaboration of a rubrics validation model.

Firstly, we managed to gather valuable information in relation to the three selected aspects of validity because we surveyed both, raters and students. We

discovered that in the elaboration of the survey, there were certain questions that provided more and less significant data for the study. This means that there are more important questions to ask at the beginning of the survey when it comes to the elaboration of a rubrics validation model. For example, how much do you know the rubric for the oral exam? or how is the rubric explained to students? are questions that we think would provide more valuable data to the study. This is relevant for a rubric validation since the evidence provided in this study indicate that lack of knowledge of a rubric has serious consequences for reliability.

Furthermore, if we take into consideration the fact that many students may not be provided with proper access to the rubric, it can be said that it is critical for any rubric validation model to take into account the extent to which reliability is ensured by providing quality access to the rubric. This could be achieved by making raters understand the importance of the rubric for the validity of the oral test so they can incorporate a detailed explanation of the rubric to the corresponding classes of the programme. As a conclusion, evidence from this study indicates that the rubric has several problems concerning construct validity, such as the ambiguities presented among the descriptors, the overlap in the description of dimensions, and the lack of alignment between the rubric and some courses of the programme.

In view of this evidence, it seems that the rubric displays deficiencies related to its nature as an adapted rubric that is not related to the actual conditions of the University and the body of students. If the rubric possesses problems in construct validity, this affects directly the reliability of the assessment since the programme is providing a defective instrument for raters to use. It seems difficult for raters to be valid and reliable with an instrument that still require work in the definition of its construct. In the same line, the evidence also shows that the aspect of reliability in this rubric is not achieved due to the inconsistency raters expressed giving more importance to one aspect instead of all aspects equally, which must be taken into account in the elaboration of a validation model for rubrics.

According to the issues mentioned in this section, even though there is no existence of a established validation model we believe our study contributes in a large extent to elaborate a validation model of rubrics taking into account the problems found in relation to the levels of construct validity, face validity, reliability and the analysis to the rubric itself.

# Chapter 5: Conclusions

The main purpose of this investigation was to analyse the validity of the oral exam rubric of the across the curriculum English Programme from USACH regarding three aspects of validity: construct validity, face validity, and reliability. We aimed at observing the main perceptions of raters and students about the oral exam rubric and the features of the actual rubric. In order to conduct this study, we proposed three research questions asking about the perceptions of raters and students about the three aspects of validity just indicated. Data was collected through surveys, and the observation of the programme of each level of English at USACH and the oral exam rubric they used. We thus identified similar and different perceptions, ideas and opinions from raters and students respecting the different aspects of validity. Simultaneously, we observed the different components of the rubric in search for validity issues in the same three dimensions.

## 5.1 Summary of results

Under RQ1.1, it can be said that evidence indicates the existence of different interpretations from the raters in relation to the application of the rubric. This affects importantly the reliability of the instrument since we can anticipate a level of inconsistency amongst raters at the moment of evaluating and grading triggered by the different perceptions that each rater might have in relation to the rubric.

Evidence also indicates that raters did not have a consistent training system required in order to use and apply the rubric correctly and successfully. This has again an important impact on the reliability of the instrument due to the fact that many features of the rubric have not been explained to the raters, and therefore, have not been internalised by them. This issue might eventually cause disparities from one rater to the other when it comes to grade the students.

Despite the previous issues, the rubric presents high levels of face validity according to the raters even though the assessment instruments present issues regarding its construct. In this context, face validity seems to be more an indicator of the unawareness of the rubric from the raters and students rather than a result of a clear knowledge and understanding of it.

This confirms the idea that the validity of an instrument of evaluation is unitary when it comes to give a general outlook of the actual functioning of, in this case, the rubric. At the same time, however, the validity of the instrument is also proven to be multi-faceted since it has various components and sources such as those examined in this study (ace validity, construct validity and reliability).

Under RQ1.2, results from students' perceptions regarding construct validity do not seem conclusive since the evidence reveals that students were not properly informed about the programme and assessment instrument and there were some students that did not know the instrument at all. Regarding reliability, students perceived that raters gave more emphasis to one dimension of the rubric over others and it was evident that some raters did not explain or show the rubric, affecting the reliability of the instrument and evaluation. Finally, the face validity of the rubric was well-evaluated by the students, in addition to the programme in general and the tools it offers. However, considering the fact that students have little access to and explanations of the rubric, this positive face validity can also be safely interpreted as an indication of lack of knowledge of the rubric.

Under RQ1.3 it can be said that the construct validity of the rubric is seriously affected by issues in all its components. In particular, there is no clear alignment between what the rubric measures and the descriptors and the contents that are taught in the courses could be observed. In addition to this, the existence of fuzzy descriptors promotes the effect, in some cases, of dimension overlap, when the boundaries between two dimensions are not clear enough and thus allow for multiple interpretations by raters. This clearly affects the validity of the construct of the rubric

in relation to the programme since its lack of accuracy can confuse both raters and students being evaluated.

The main findings of this study are related to the three aspects of validity investigated: construct validity, face validity and reliability. Observations regarding construct validity of the rubrics include the misalignment of the rubric with the programme, the lack of dimensions the stakeholders declared such as Pronunciation and Authenticity, and the ambiguity of its descriptors. In the second place, the findings also show that there is a satisfactory face validity of the assessment instrument. However, this face validity must be interpreted cautiously due to the evidence that indicate that raters do not know much of the rubric and students are not acquainted with it either. In this way, this study illustrates how construct validity and face validity are independent from each other. Following this, findings have revealed that raters have no training to use the rubric and most of them did not participate in the design of the rubric, allowing for different interpretations of the same rubric. Besides, there is a lack of procedures for ensuring the reliability of the rubric. The flaws detected in the rubric regarding its construct validity make the rubric a defective instrument by limiting the consistent use of it by raters. This shows how critical is reliability for the validity of the rubric.

The factors affecting the validity of the rubric can be explained in part by the way in which the rubric was adapted from the original Cambridge scale. It has been argued in this thesis that rubrics must, in general, respond to the needs and conditions of the contexts in which they are applied. It thus seems necessary to take into account the reality of the educational landscape of the country, the reality of the University and the body of students when revising the design of the rubric observed in this study.

## 5.2 Limitations to the study and suggestions for further research

The conclusions indicated in section 6.1 must be considered against some important limitations to the study. In the first place, we only surveyed students from three different programs at USACH, so our results and conclusions are not representative of the the entire Blended Learning programme. Perceptions observed in this study are nevertheless consistent enough to assume that they may be confirmed in a broader study that considered more students from other programmes. .

One of the main conclusions of the study was that the high face validity of the rubric had to be interpreted cautiously as participants also demonstrated a low degree of knowledge of the rubric. In the study reported in this thesis, we did not investigate the possible factors that could explain these results. Further research should address this issue explicitly by considering the observation of variables that could explain a positive evaluation of the rubric associated to a low knowledge of it. Such variables may include, for example, lack of motivation to respond to the surveys or the lack of importance of the rubric in the expectations of students to pass their courses of English.

We believe that the result of the combination of analyzing the rubric and stakeholders' perceptions of the rubric was useful and significant for identifying the issues of the implementation of the rubric for the oral exam and also, for describing the ways in which stakeholders interact with the oral exam rubric. Further research should be carried out in order to validate and systematise the procedures explored in our study.

This study contributes to the current state of knowledge regarding the use of rubrics in higher education. The analysis of the rubric by itself and the analysis of the stakeholders' perceptions allowed us to observe the interaction of construct validity and face validity and the fundamental role that reliability plays in the validity of rubrics.

The evidence provided in this study also highlights the critical need for implementing procedures for ensuring reliability in standardised contexts. Such procedures need to be aimed especially at making raters aware and confident in the use of the rubric, giving adequate access to the rubric to students and, in general, promoting the use of rubrics in the teaching process and not only as a tool for scoring.

In general, we hope that the evidence from this study contributes to a much-needed discussion about the validity of rubrics as a topic which is related, but different, from the validity of tests. In this way, our study will be contributing to develop a more comprehensive view of validity in language assessment processes.

# REFERENCES


Andrade, H. 2000. Using rubrics to promote thinking and learning. Educational Leadership 57, no. 5: 13–18.

Reddy, Y., & Andrade, H. (2010). A review of rubric use in higher education. Assessment & Evaluation In Higher Education, 35(4), 435-448. doi: 10.1080/02602930902862859

Arther, J., & McTighe, J. (2001). Scoring Rubrics in the Classroom. Thousand Oaks: Corwin Press.

Bachman, L. (1990). Fundamental considerations in language testing. Oxford: Oxford University Press.

Bachman, Lyle. (2013). Ongoing Challenges in Language Assessment. 10.1002/9781118411360.wbcla128.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. Psychological Review, 111, 1061–1071.

Brown, G., Bull, J. and Pendlebury, M. (1997) Assessing Student Learning in Higher Education. Routledge, London.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. Language testing, 20(1).

Brown, H. D. (2004). Language assessment: Principles and classroom practices. Longman Pub Group

Cumming, A. (2009). LANGUAGE ASSESSMENT IN EDUCATION: TESTS, CURRICULA, AND TEACHING. Annual Review of Applied Linguistics, 29, 90-100. doi:10.1017/S0267190509090084

Gronlund, N. (1998). Assessment of Student Achievement. Sixth Edition. Retrieved from https://eric.ed.gov/?id=ED417221

Green, R., and M. Bowser. 2006. Observations from the field: Sharing a literature review rubric. Journal of Library Administration 45, nos. 1–2: 185–202.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences [Ebook]. Malmo, Sweden: School of Teacher Education, Malmo University,. Retrieved from http://www.sciencedirect.com

Luoma, S. (2004). Assessing speaking. New York: Cambridge University Press.

Mertler, Craig A. (2001). Designing scoring rubrics for your classroom. Practical Assessment, Research & Evaluation, 7

Messick, S. (1989). Validity. In R. L. Lin (Ed.), Educational measurement (3rd ed., p. 13-103). New York: Macmillan.

Messick, S. (1996). Validity and washback in language testing. Princeton, N.J.: Educational Testing Service.

Messick, S. (1996). Validity of performance assessments. In G. Phillips (Ed.), Technical issues in large-scale performance assessment (pp. 1–18). Washington, DC: National Center for Education Statistics

MINEDUC. (2009) Objetivos Fundamentales y Contenidos Mínimos Obligatorios de la Educación Básica y Media. Retrieved from https://www.agenciaeducacion.cl/wp-content/uploads/2013/02/Marco-Curricular-y-Actualizacion-2009-I-a-IV-Medio.pdf

Moskal B.M., & Leydens J.A. (2000). Scoring rubric development: validity and reliability. Practical Assessment, Research and Evaluation.

Moni, R.W., E. Beswick, and K.B. Moni. 2005. Using student feedback to construct an assessment rubric for a concept map in physiology. Advances in Physiology Education 29: 197–203.

Mousavi, S.A. (2002). An encyclopedic dictionary language testing: (Third edition). Taipei: Tung Hua Publications.

McMillan, J. H. (2004). Educational research: Fundamentals for the consumer. Boston: Pearson Education Inc..

Payne, D.A. 2003. Applied educational assessment. 2nd ed. Belmont, CA: Wadsworth/ Thomson Learning

Stevens, D., Levi, A. and Walvoord, B. (2013). Introduction to rubrics. Sterling, Virginia: Stylus Publishing.

# APPENDIX

The following link contains the Appendix for this study:

https://drive.google.com/open?id=1Jt4prB4DjzwrUlkkquk8VtwglMFqBibv